

Solanum lycopersicum (tomato) Transcriptome Sequencing Report

June 2022



Project Information

Client Name	Victor Flors
Company/Institution	Universitat Jaume I
Order Number	HN00171344
Species	<i>Solanum lycopersicum</i> (tomato)
Reference	SL3.0
Annotation	GCF_000188115.4
Type of Read	Paired-ends
Read Length	151
Number of Samples	24
Library Kit	TruSeq Stranded mRNA Library Prep Kit
Library Protocol	TruSeq Stranded mRNA Reference Guide # 1000000040498 v00
Type of Sequencer	Illumina platform

Project Results Summary

In this study, *Solanum lycopersicum* (tomato) whole transcriptome sequencing was performed in order to examine the different gene expression profiles, and to perform gene annotation on set of useful genes based on gene ontology pathway information.

Analyses were successfully performed on all 24 paired-ends samples. Figure 1 shows the throughput of raw data and trimmed data. Figure 2 shows the Q30 percentage (% of bases with quality over phred score 30) of each sample's raw and trimmed data.

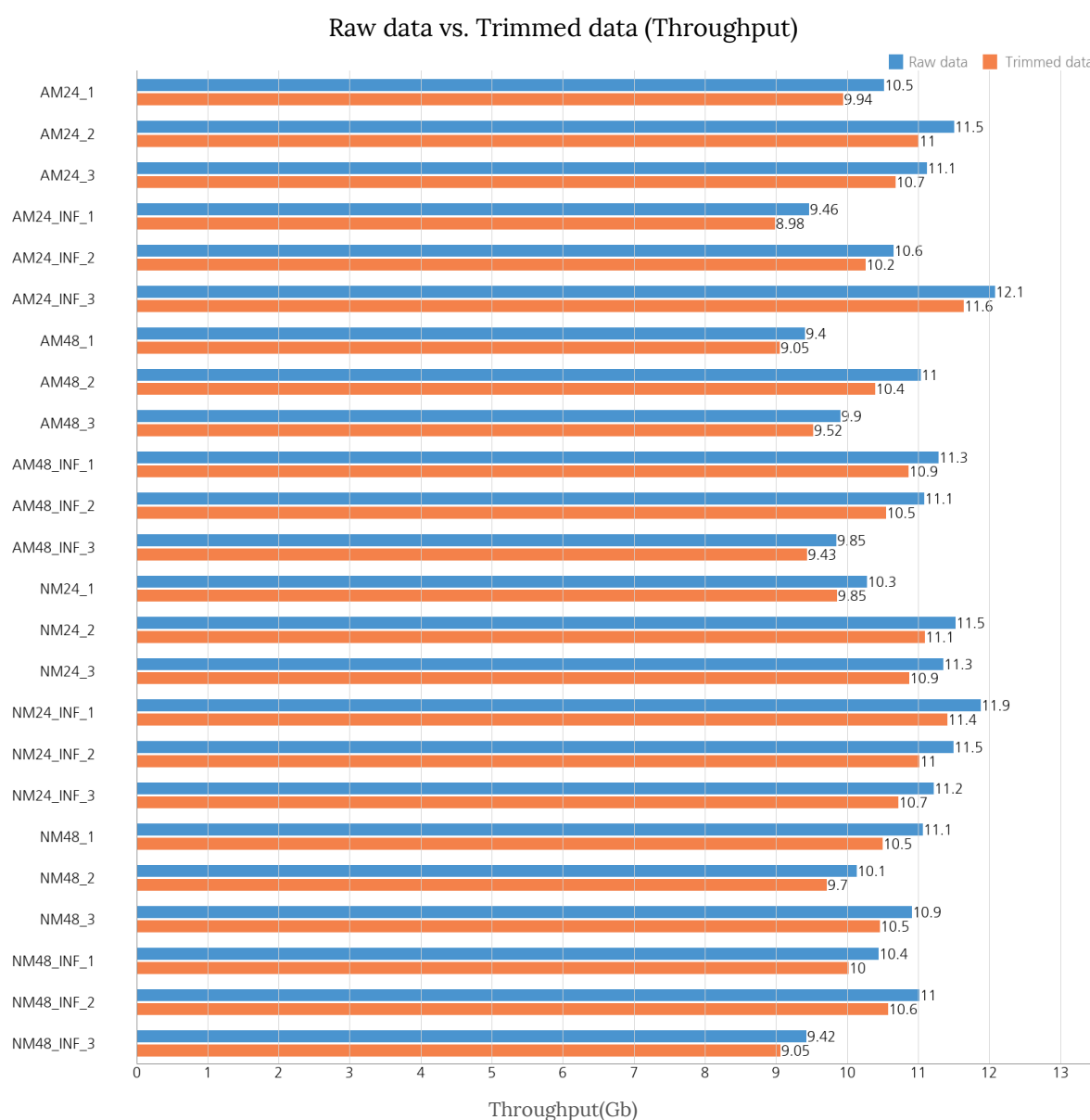


Figure 1. Throughput output of Raw and Trimmed data

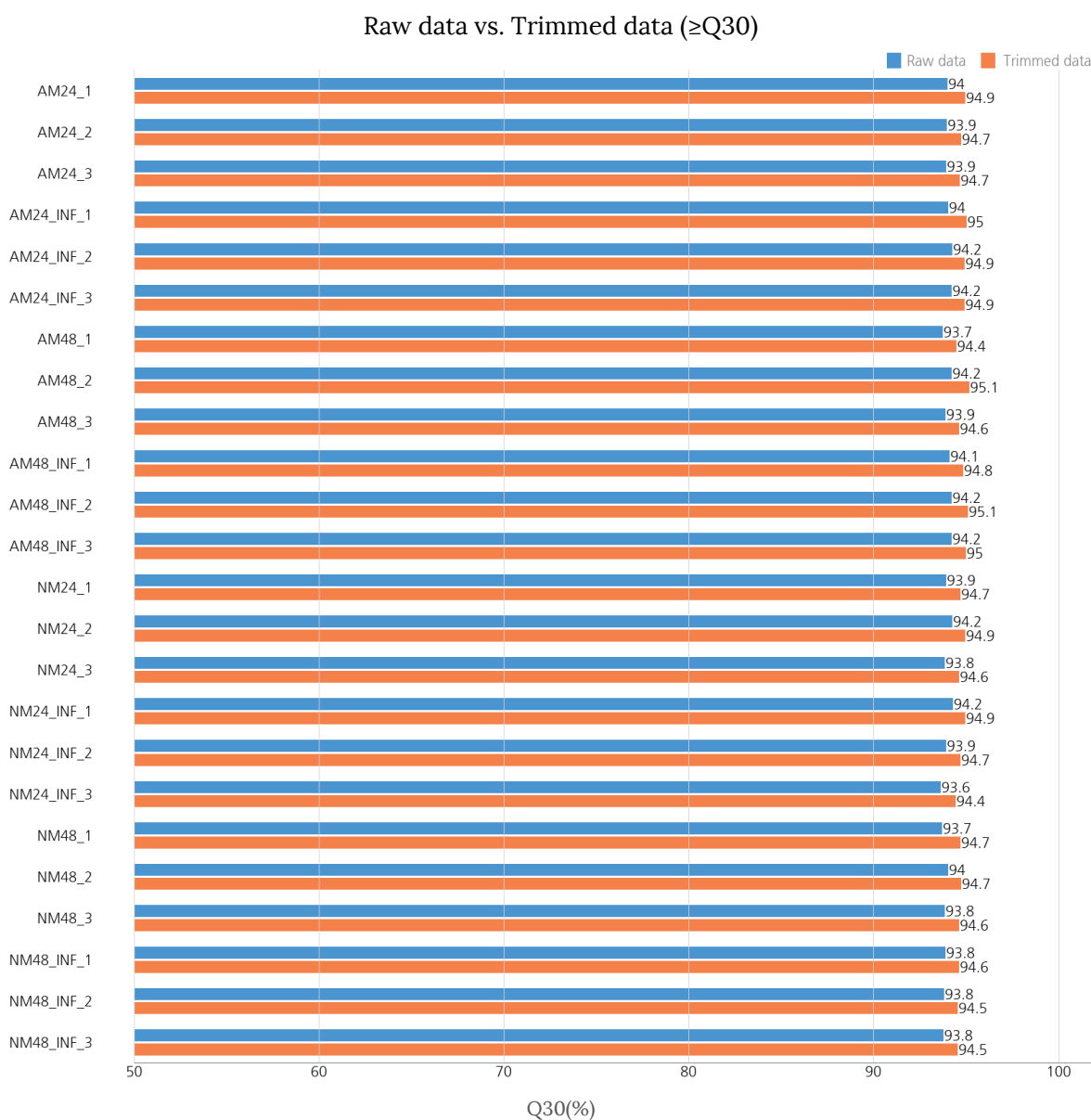


Figure 2. Q30 score of Raw and Trimmed data

Trimmed reads are mapped to reference genome with HISAT2. Figure 3 shows the overall read mapping ratio, the ratio of mapped reads to trimmed reads.

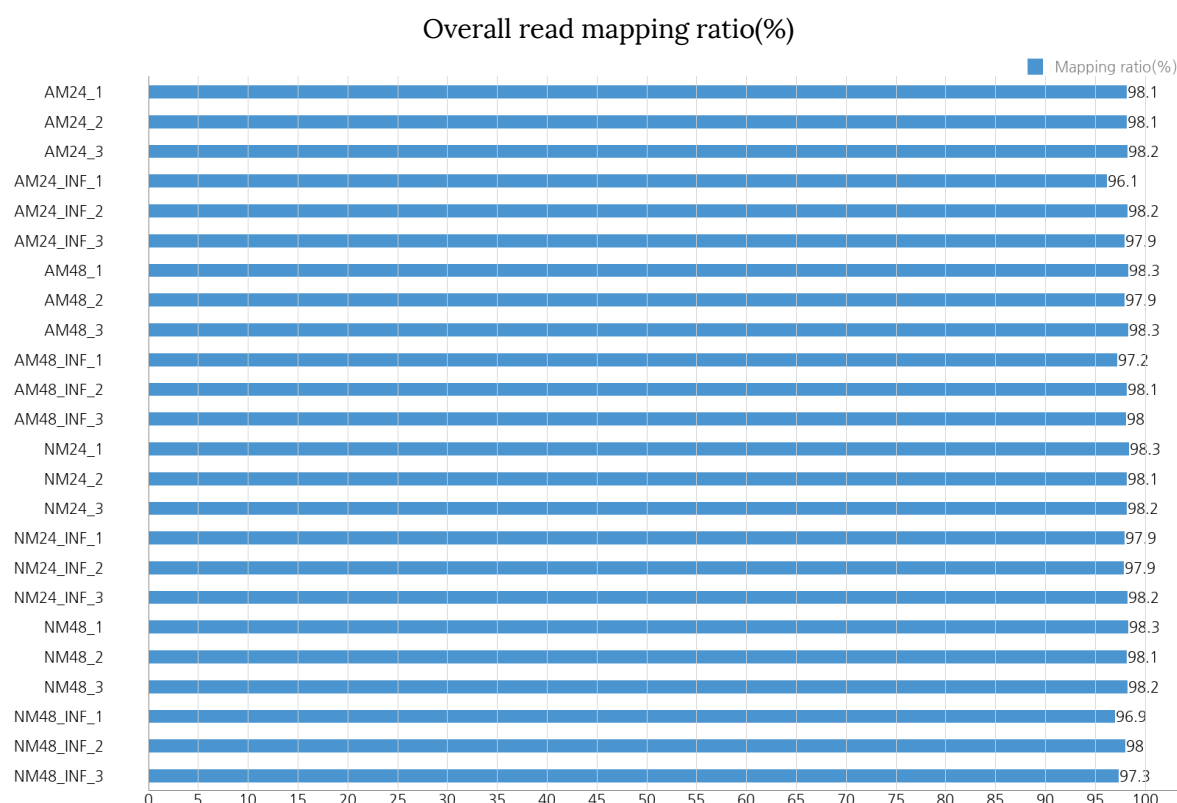


Figure 3. Overall read mapping ratio(%)

After the read mapping, Stringtie was used for transcript assembly. Expression profile was calculated for each sample and transcript/gene as read count, FPKM (Fragment per Kilobase of transcript per Million mapped reads) and TPM (Transcripts Per Kilobase Million).

DEG (Differentially Expressed Genes) analysis was performed on 8 comparisons pairs as requested using DESeq2. The results showed 4,017 genes which satisfied $|fc| \geq 2$ & $nbinomWaldTest$ raw $p\text{-value} < 0.05$ conditions in at least one of comparison pairs.

Figure 4 shows the result of hierarchical clustering (distance metric= Euclidean distance, linkage method= complete) analysis. It graphically represents the similarity of expression patterns between samples and genes.

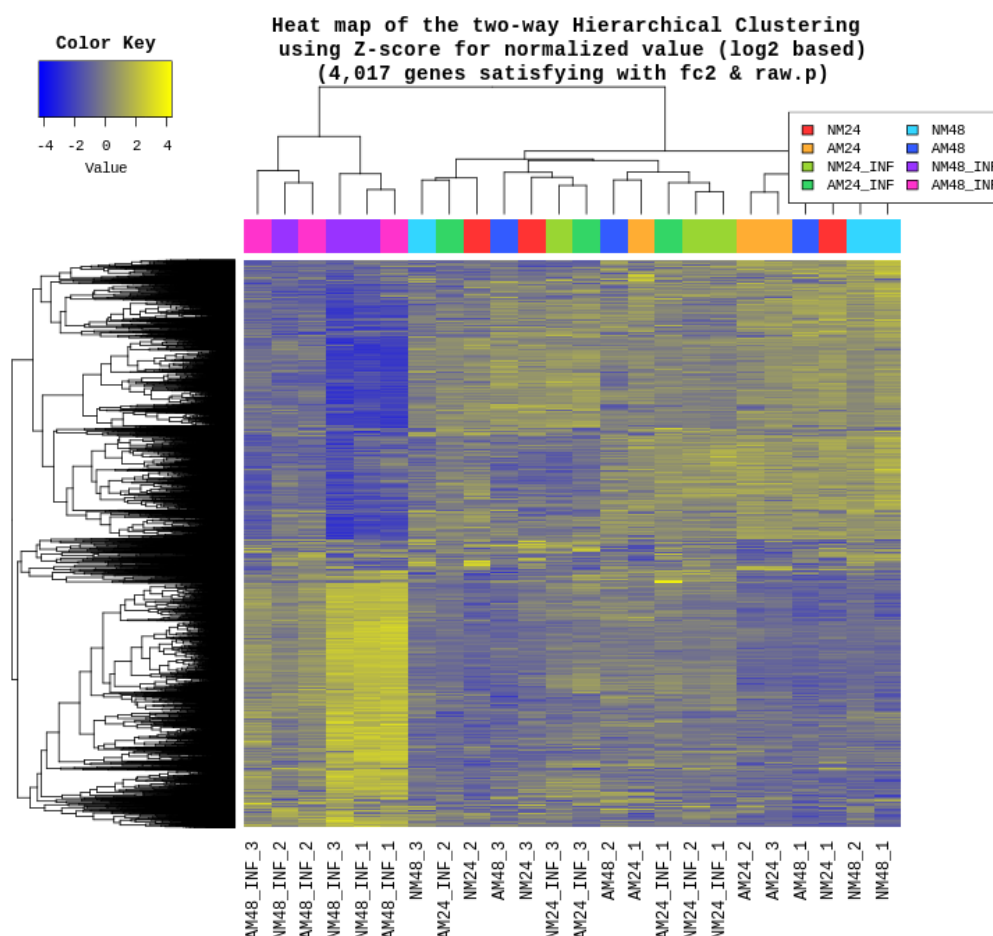


Figure 4. Heatmap for DEG list

DEG list was further analyzed with gProfiler (<https://biit.cs.ut.ee/gprofiler/orth>) for gene set enrichment analysis per biological process (BP), cellular component (CC) and molecular function (MF). The Figure 5, 6 and 7 show the significant gene set by each category.

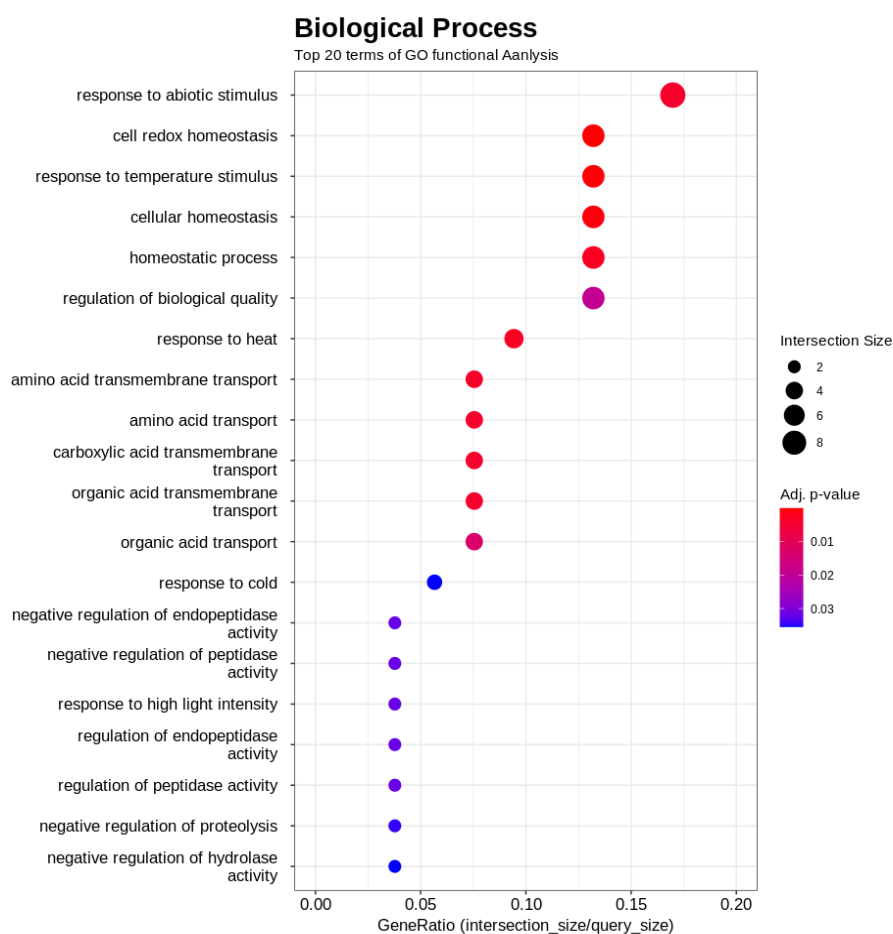


Figure 5. Gene Ontology terms related to Biological Process

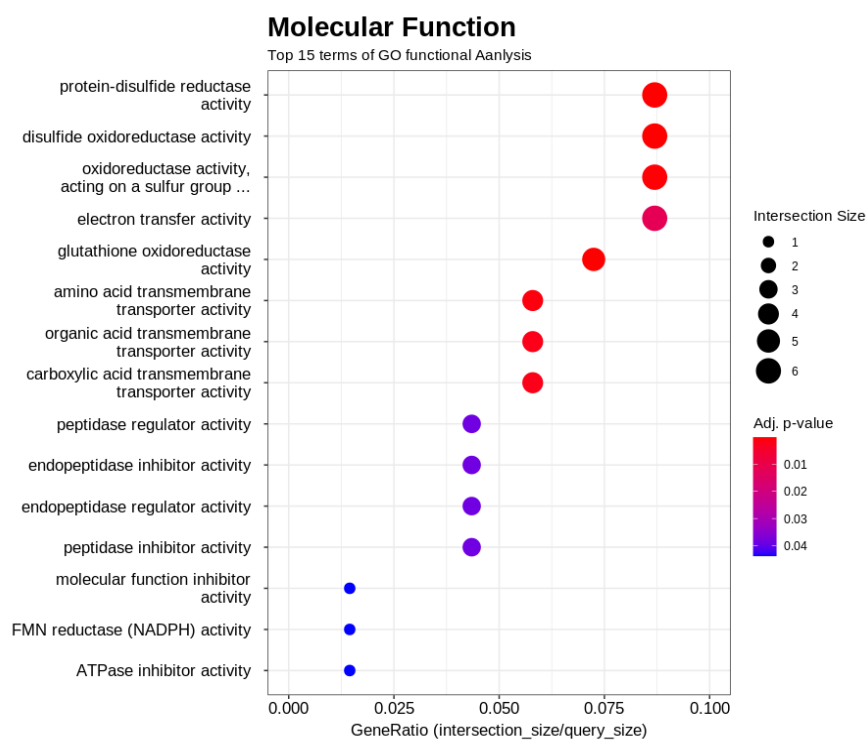


Figure 6. Gene Ontology Terms related to Molecular Function

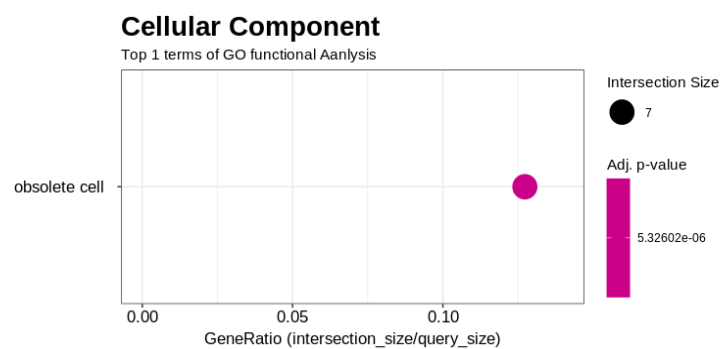


Figure 7. Gene Ontology Terms related to Cellular Component

Table of Contents

Project Information	02
Project Results Summary	03
1. Experimental Methods and Workflow	11
2. Analysis Methods and Workflow	12
3. Summary of Data Production	13
3. 1. Raw Data Statistics	13
3. 2. Average Base Quality at Each Cycle	15
3. 3. Trimming Data Statistics	16
3. 4. Average Base Quality at Each Cycle after Trimming	18
4. Reference Mapping and Assembly Results	19
4. 1. Mapping Data Statistics	19
4. 2. Expression Profiling	21
5. Differentially Expressed Gene Analysis Results	23
5. 1. Data Analysis Quality Check and Preprocessing	23
5. 2. Differentially Expressed Gene Analysis Workflow	29
5. 3. Significant Gene Results	31
5. 4. GO Enrichment Analysis	37
6. Data Download Information	43
6. 1. Raw Data	43
7. Appendix	47
7. 1. Phred Quality Score Chart	47
7. 2. Programs used in Analysis	48
7. 3. References	49

1. Experimental Methods and Workflow

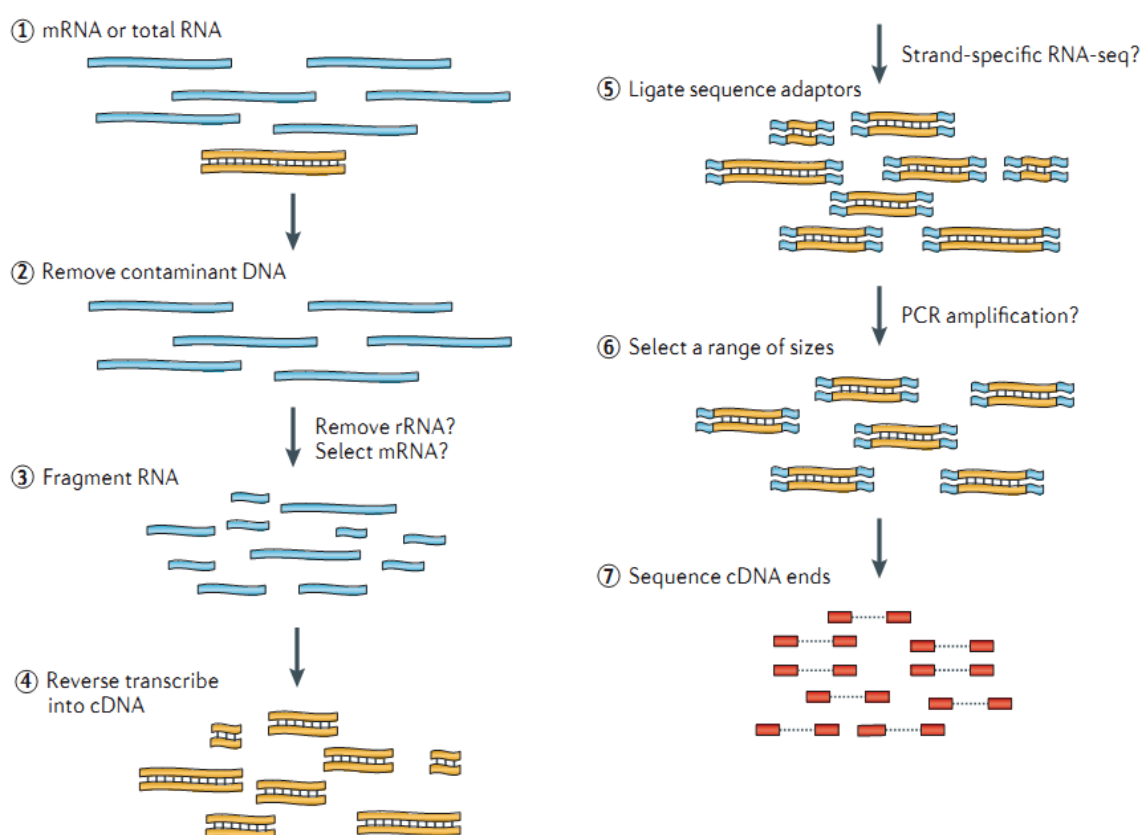


Figure 8. RNA Sequencing Experiment Workflow

REFERENCE ♦ Nat Rev Genet. 2011 Sep 7;12(10):671-82

- 1) Isolate the Total RNA from Sample of interest (Cell or Tissue).
- 2) Eliminate DNA contamination using DNase.
- 3) Choose an appropriate kit for library prep process depending on the types of RNA. For mRNA with poly-A tail, use mRNA purification kit; for non-coding RNAs, such as lincRNA, use ribo-zero RNA removal Kit to purify RNA of interest.
- 4) Randomly fragment purified RNA for short read sequencing.
- 5) Reverse transcribe fragmented RNA into cDNA.
- 6) Ligate adapters onto both ends of the cDNA fragments.
- 7) After amplifying fragments using PCR, select fragments with insert sizes between 200-400 bp. For paired-end sequencing, both ends of the cDNA is sequenced by the read length.

2. Analysis Methods and Workflow



Figure 9. Analysis Workflow

- 1) Analyze the quality control of the sequenced raw reads. Overall reads' quality, total bases, total reads, GC (%) and basic statistics are calculated.
- 2) In order to reduce biases in analysis, artifacts such as low quality reads, adaptor sequence, contaminant DNA, or PCR duplicates are removed.
- 3) Trimmed reads are mapped to reference genome with HISAT2, splice-aware aligner.
- 4) Transcript is assembled by StringTie with aligned reads.
- 5) Expression profiles are represented as read count and normalization values which are calculated based on transcript length and depth of coverage. Normalization values are provided as FPKM (Fragments Per Kilobase of transcript per Million Mapped reads) / RPKM (Reads Per Kilobase of transcript per Million mapped reads) and TPM (Transcripts Per Kilobase Million).
- 6) In groups with different conditions, genes or transcripts that express differentially are filtered out through statistical hypothesis testing.

3. Summary of Data Production

3.1. Raw Data Statistics

(Refer to Path: result_RNAseq/Analysis_statistics/rawData/raw_throughput.stats)

The total number of bases, reads, GC (%), Q20 (%), Q30 (%) are calculated for 24 samples. For example, in AM24_1, 69,633,566 reads are produced, and total read bases are 10.5Gbp. The GC content (%) is 43.77% and Q30 is 93.98%.

Table 1. Raw data stats

Index	Sample id	Total read bases*	Total reads	GC (%)	Q20 (%)	Q30 (%)
1	AM24_1	10,514,668,466	69,633,566	43.77	97.91	93.98
2	AM24_2	11,498,797,074	76,150,974	43.61	97.97	93.94
3	AM24_3	11,115,521,626	73,612,726	43.53	97.96	93.89
4	AM24_INF_1	9,460,251,774	62,650,674	44.12	97.87	94.01
5	AM24_INF_2	10,649,050,614	70,523,514	43.54	98.13	94.23
6	AM24_INF_3	12,070,093,494	79,934,394	43.65	98.10	94.20
7	AM48_1	9,399,474,576	62,248,176	43.72	97.92	93.70
8	AM48_2	11,032,402,770	73,062,270	43.76	98.01	94.19
9	AM48_3	9,898,831,878	65,555,178	43.69	97.96	93.86
10	AM48_INF_1	11,282,027,212	74,715,412	42.94	98.04	94.08
11	AM48_INF_2	11,074,301,646	73,339,746	43.52	98.02	94.21
12	AM48_INF_3	9,845,381,804	65,201,204	43.12	98.07	94.18
13	NM24_1	10,276,414,324	68,055,724	43.68	97.97	93.89
14	NM24_2	11,514,610,700	76,255,700	43.50	98.10	94.23
15	NM24_3	11,346,004,100	75,139,100	43.56	97.96	93.84
16	NM24_INF_1	11,871,402,862	78,618,562	43.50	98.12	94.25
17	NM24_INF_2	11,493,475,230	76,115,730	43.59	97.95	93.90
18	NM24_INF_3	11,208,001,274	74,225,174	43.79	97.84	93.60
19	NM48_1	11,059,118,596	73,239,196	43.64	97.79	93.68
20	NM48_2	10,127,611,676	67,070,276	43.55	98.02	94.00
21	NM48_3	10,900,966,934	72,191,834	43.65	97.93	93.84
22	NM48_INF_1	10,437,650,312	69,123,512	43.08	97.95	93.85
23	NM48_INF_2	11,012,693,646	72,931,746	43.52	97.93	93.78
24	NM48_INF_3	9,418,555,238	62,374,538	43.12	97.91	93.75

(* Total read bases = Total reads x Read length)

- Total read bases: Total number of bases sequenced
- Total reads: Total number of reads

- GC (%): GC content
- Q20 (%): Ratio of bases that have phred quality score greater than or equal to 20
- Q30 (%): Ratio of bases that have phred quality score greater than or equal to 30

3. 2. Average Base Quality at Each Cycle

(Refer to Path: Analysis_statistics/rawData/A_fastqc/)

The quality of produced data is determined by the phred quality score at each cycle. Box plot containing the average quality at each cycle is created with FastQC.

The x-axis shows number of cycles and y-axis shows phred quality score. Phred quality score 20 means 99% accuracy and reads over score of 20 are accepted as good quality.

LINK <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>

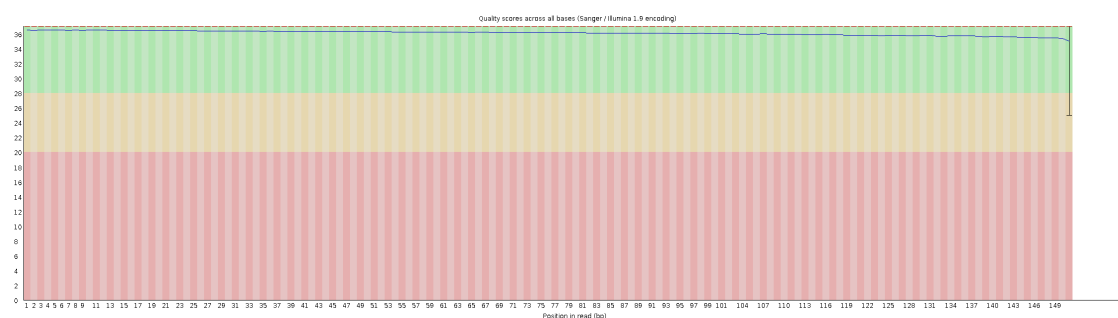


Figure 10. Read quality at each cycle of AM24_1 (read1)

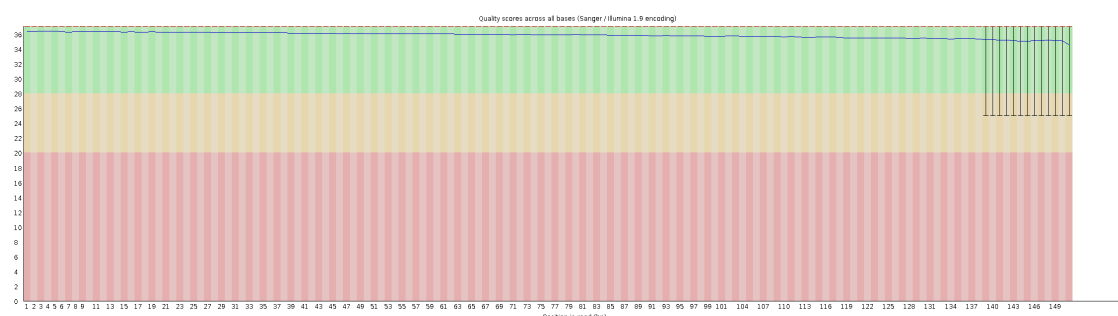


Figure 11. Read quality at each cycle of AM24_1 (read2)

- Yellow box: Interquartile range (25-75%) of phred score at each cycle
- Red line: Median of phred score at each cycle
- Blue line: Average of phred score at each cycle
- Green background: Good quality
- Orange background: Acceptable quality
- Red background: Bad quality

3. 3. Trimming Data Statistics

(Refer to Path: result_RNAseq/Analysis_statistics/trimmedData/trim_throughput.stats)

Trimmomatic program is used to remove adapter sequences and bases with base quality lower than three from the ends. Also using sliding window method, bases of reads that does not qualify for window size 4, and mean quality 15 are trimmed. Afterwards, reads with length shorter than 36bp are dropped to produce trimmed data.

Table 2. Trimming Data Stats

Index	Sample id	Total read bases	Total reads	GC(%)	Q20(%)	Q30(%)
1	AM24_1	9,937,785,645	68,515,368	43.58	98.59	94.93
2	AM24_2	11,004,304,710	75,209,586	43.46	98.51	94.72
3	AM24_3	10,675,513,767	72,725,998	43.41	98.49	94.65
4	AM24_INF_1	8,978,912,627	61,536,874	44.01	98.60	95.02
5	AM24_INF_2	10,249,042,711	69,793,114	43.40	98.60	94.90
6	AM24_INF_3	11,632,715,094	79,028,544	43.52	98.58	94.89
7	AM48_1	9,048,607,265	61,513,808	43.60	98.43	94.44
8	AM48_2	10,384,818,555	71,894,284	43.54	98.67	95.14
9	AM48_3	9,515,881,401	64,765,044	43.55	98.47	94.60
10	AM48_INF_1	10,853,810,547	73,880,066	42.76	98.54	94.81
11	AM48_INF_2	10,537,526,805	72,209,984	43.37	98.64	95.09
12	AM48_INF_3	9,428,435,507	64,422,592	42.99	98.59	94.96
13	NM24_1	9,851,689,635	67,203,078	43.54	98.50	94.66
14	NM24_2	11,085,567,572	75,408,094	43.36	98.59	94.93
15	NM24_3	10,869,247,173	74,252,810	43.38	98.47	94.58
16	NM24_INF_1	11,406,502,582	77,764,946	43.34	98.60	94.94
17	NM24_INF_2	11,007,323,665	75,136,428	43.43	98.50	94.69
18	NM24_INF_3	10,714,778,602	73,267,574	43.59	98.40	94.40
19	NM48_1	10,489,957,374	71,973,628	43.47	98.49	94.66
20	NM48_2	9,702,278,750	66,324,262	43.39	98.52	94.72
21	NM48_3	10,458,830,521	71,262,260	43.49	98.47	94.61
22	NM48_INF_1	10,012,759,000	68,295,922	42.90	98.48	94.59
23	NM48_INF_2	10,574,371,731	72,023,948	43.38	98.46	94.54
24	NM48_INF_3	9,053,738,363	61,592,422	42.97	98.44	94.52

- Total read bases: Total number of read bases after trimming
- Total reads: Total number of reads after trimming
- GC (%): GC Content
- Q20 (%): Ratio of bases that have phred quality score greater than or equal to 20

- Q30 (%): Ratio of bases that have phred quality score greater than or equal to 30

3. 4. Average Base Quality at Each Cycle after Trimming

(Refer to Path: result_RNAseq/Analysis_statistics/trimmedData/A_fastqc/)

Figure 12 and 13 show average base quality at each cycle after trimming.

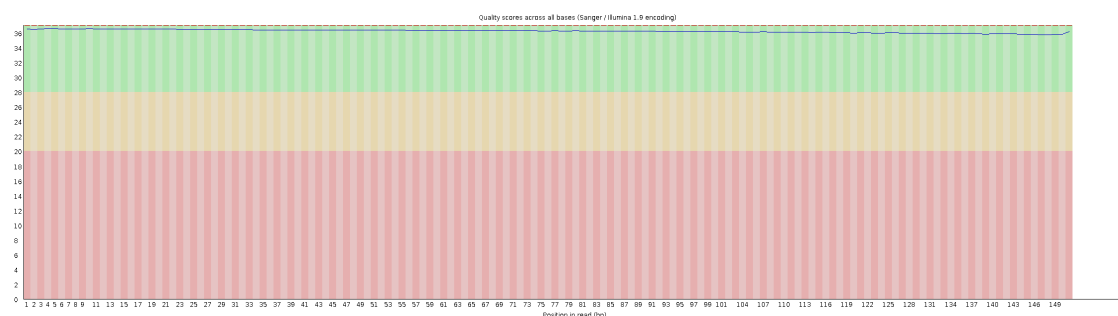


Figure 12. Average base quality of AM24_1 (read1) at each cycle after trimming

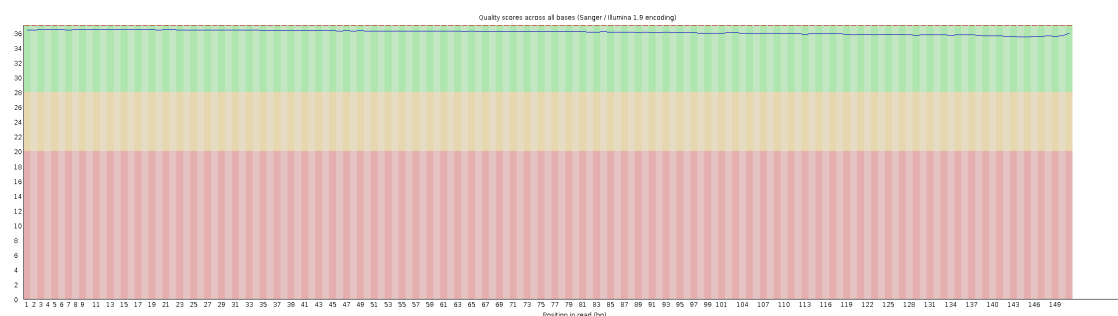


Figure 13. Average base quality of AM24_1 (read2) at each cycle after trimming

- Yellow box: Interquartile range (25–75%) of phred score at each cycle
- Red line: Median of phred score at each cycle
- Blue line: Average of phred score at each cycle
- Green background: Good quality
- Orange background: Acceptable quality
- Red background: Bad quality

4. Reference Mapping and Assembly Results

4.1. Mapping Data Statistics

(Refer to Path: result_RNAseq/Analysis_statistics/mapping.hisat.stats)

In order to map cDNA fragments obtained from RNA sequencing, SL3.0 was used as a reference genome. Table 3 shows the statistic obtained from HISAT2, which is known to handle spliced read mapping through Bowtie2 aligner. You can check number of processed reads, mapped reads.

Table 3. Mapped Data Stats

Sample ID	# of processed reads	# of mapped reads (%)	# of unmapped reads (%)
AM24_1	68,515,368	67,221,524 (98.11%)	1,293,844 (1.89%)
AM24_2	75,209,586	73,793,586 (98.12%)	1,416,000 (1.88%)
AM24_3	72,725,998	71,440,753 (98.23%)	1,285,245 (1.77%)
AM24_INF_1	61,536,874	59,139,001 (96.1%)	2,397,873 (3.9%)
AM24_INF_2	69,793,114	68,511,173 (98.16%)	1,281,941 (1.84%)
AM24_INF_3	79,028,544	77,365,548 (97.9%)	1,662,996 (2.1%)
AM48_1	61,513,808	60,449,869 (98.27%)	1,063,939 (1.73%)
AM48_2	71,894,284	70,368,829 (97.88%)	1,525,455 (2.12%)
AM48_3	64,765,044	63,659,177 (98.29%)	1,105,867 (1.71%)
AM48_INF_1	73,880,066	71,781,388 (97.16%)	2,098,678 (2.84%)
AM48_INF_2	72,209,984	70,829,282 (98.09%)	1,380,702 (1.91%)
AM48_INF_3	64,422,592	63,144,798 (98.02%)	1,277,794 (1.98%)
NM24_1	67,203,078	66,076,065 (98.32%)	1,127,013 (1.68%)

NM24_2	75,408,094	73,987,315 (98.12%)	1,420,779 (1.88%)
NM24_3	74,252,810	72,879,910 (98.15%)	1,372,900 (1.85%)
NM24_INF_1	77,764,946	76,157,924 (97.93%)	1,607,022 (2.07%)
NM24_INF_2	75,136,428	73,526,847 (97.86%)	1,609,581 (2.14%)
NM24_INF_3	73,267,574	71,931,814 (98.18%)	1,335,760 (1.82%)
NM48_1	71,973,628	70,745,453 (98.29%)	1,228,175 (1.71%)
NM48_2	66,324,262	65,063,181 (98.1%)	1,261,081 (1.9%)
NM48_3	71,262,260	69,993,762 (98.22%)	1,268,498 (1.78%)
NM48_INF_1	68,295,922	66,186,954 (96.91%)	2,108,968 (3.09%)
NM48_INF_2	72,023,948	70,548,727 (97.95%)	1,475,221 (2.05%)
NM48_INF_3	61,592,422	59,950,529 (97.33%)	1,641,893 (2.67%)

- Processed reads: Number of cleaned reads after trimming
- Mapped reads: Number of reads mapped to reference
- Unmapped reads: Number of reads that failed to align

4. 2. Expression Profiling

Known genes and transcripts are assembled with StringTie based on reference genome model.

After assembly, the abundance of gene/transcript is calculated in the read count and normalized values as FPKM (Fragments Per Kilobase of transcript per Million mapped reads) and TPM (Transcripts Per Kilobase Million) for a sample.

4. 2. 1. Known Transcripts Expression Level

(Refer to Path: result_RNAseq_excel/Expression_profile/StringTie/Expression_Profile.SL3.0.transcript.xlsx)

Table 4 is an example of known transcript expression level per sample in expression value. This result is obtained by -e option of StringTie does not consider novel transcript assembly.

Table 4. Known transcripts Expression Level (example)

Transcript_ID	Gene_ID	Gene Symbol	Description	Transcript_Locus	Transcript Length	AM Read_Count	BM Read_Count	AM_FPKM	BM_FPKM	AM_TPM	BM_TPM
NM_130786	1	A1BG	alpha-1-B glycoprotein	chr19:58345183-58353492	3382	88	163	0.432396	0.678319	0.947053	1.504474
NR_040112	3	A2MP1	alpha-2-macroglobulin pseudog	chr12:9228533-9234207	1201	0	0	0	0	0	0
XM_017013947	9	NAT1	N-acetyltransferase 1, transcrip	chr8:18170419-18223689	2704	0	21	0	0.108737	0	0.241173
NM_001291962	9	NAT1	N-acetyltransferase 1, transcrip	chr8:18170467-18223689	2122	0	0	0	0	0	0
NM_000015	10	NAT2	N-acetyltransferase 2	chr8:18391282-18401218	1285	0	0	0	0	0	0
NM_001085	12	SERPINA3	serpin family A member 3	chr14:94612377-94624053	1590	8	75	0.084216	0.664787	0.184454	1.474461
XM_005247104	13	AADAC	arylacetamide deacetylase, tra	chr3:151814008-151828488	1620	0	12	0	0.102866	0	0.228152
NM_001086	13	AADAC	arylacetamide deacetylase	chr3:151814116-151828488	1563	108	108	1.152579	0.971041	2.524427	2.153715
XM_024452712	14	AAMP	angio associated migratory cell	chr2:218264127-218270181	2002	106	101	0.879142	0.710738	1.925533	1.576378
NM_001302545	14	AAMP	angio associated migratory cell	chr2:218264129-218270137	1763	1621	1797	15.408498	14.424821	33.74635	31.99344
NM_001087	14	AAMP	angio associated migratory cell	chr2:218264129-218270137	1760	9332	10212	88.854179	82.119453	194.6122	182.1363
NM_001166579	15	AANAT	aralkylamine N-acetyltransferas	chr17:76453351-76470117	1913	2	8	0.010678	0.052728	0.023387	0.116948
XM_017024259	15	AANAT	aralkylamine N-acetyltransferas	chr17:76465946-76470797	4252	4	11	0.013221	0.03452	0.028958	0.076564
NR_110548	15	AANAT	aralkylamine N-acetyltransferas	chr17:76467548-76470117	1082	0	0	0	0	0	0
NM_001088	15	AANAT	aralkylamine N-acetyltransferas	chr17:76467603-76470117	971	0	0	0	0	0	0
XR_933220	16	AARS	alanyl-tRNA synthetase, transcr	chr16:70252295-70289509	3258	90	160	0.461517	0.694592	1.010834	1.540566
NM_001605	16	AARS	alanyl-tRNA synthetase	chr16:70252394-70289509	3344	22367	68204	112.089745	288.669189	245.5037	640.2521

- Transcript_ID: Splicing variant (isoform/transcript)
- Gene_ID: Gene ID
- Gene_Symbol: Symbol of gene
- Gene_Description: Description of gene
- Transcript_Locus: Transcript locus
- Transcript_Length: Transcript length
- [Sample Name]_Read_Count: Read count of a sample
- [Sample Name]_FPKM: FPKM normalized value of a sample
- [Sample Name]_TPM: TPM normalized value of a sample

4. 2. 2. Known Genes Expression Level

(Refer to Path: result_RNAseq_excel/Expression_profile/StringTie/
Expression_Profile.SL3.0.gene.xlsx)

Table 5 is an example of known gene expression level per sample in expression value. This result is obtained by -e option of StringTie does not consider novel transcript assembly.

Table 5. Known genes Expression Level (example)

Gene_ID	Transcript_ID	Gene Symbol	Description	AM Read_Count	BM Read_Count	AM_FPKM	BM_FPKM	AM_TPM	BM_TPM
1	NM_130786	A1BG	alpha-1-B glycoprotein	88	163	0.432396	0.678319	0.947053	1.504474
2	NM_000014.NM_001347423	A2M	alpha-2-macroglobulin	0	0	0	0	0	0
3	NR_040112	A2MP1	alpha-2-macroglobulin pseudogene	0	0	0	0	0	0
9	NM_000682.NM_001160170	NAT1	N-acetyltransferase 1	288	217	2.411185	1.490984	5.281078	3.306918
10	NM_000015.XM_017012938	NAT2	N-acetyltransferase 2	10	6	0.097138	0.050729	0.212756	0.112513
12	NM_001085	SERPINA3	serpin family A member 3	8	75	0.084216	0.664787	0.184454	1.474461
13	NM_001086.XM_005247104	AADAC	arylacetamide deacetylase	108	120	1.152579	1.073907	2.524427	2.381867
14	NM_001087.NM_001302545	AAMP	angio associated migratory cell prot	11059	12110	105.141819	97.255012	230.2861	215.7062
15	NM_001088.NM_001166579	AANAT	aralkylamine N-acetyltransferase	6	19	0.023899	0.087248	0.052345	0.193512
16	NM_001605.XR_933220	AARS	alanyl-tRNA synthetase	22457	68364	112.551262	289.363781	246.5145	641.7927
18	NM_000683.NM_001127448	ABAT	4-aminobutyrate aminotransferase	327	175	1.143824	0.441216	2.505251	0.978593
19	NM_005502.XM_005251773	ABCA1	ATP binding cassette subfamily A n	1496	2718	2.403716	3.695532	5.264719	8.196482
20	NM_001606.NM_212533.XM	ABCA2	ATP binding cassette subfamily A n	2500	3986	5.218521	6.986245	11.42982	15.4951
21	NM_001089	ABCA3	ATP binding cassette subfamily A n	2214	4876	5.619098	10.452255	12.30719	23.18251
22	NM_001271696.NM_0012716	ABCB7	ATP binding cassette subfamily B n	2618	1974	9.550061	6.097788	20.91695	13.52455
23	NM_001025091.NM_001090	ABCF1	ATP binding cassette subfamily F n	11449	11921	56.366045	49.563715	123.4553	109.9295
24	NM_000350	ABCA4	ATP binding cassette subfamily A n	62	139	0.140036	0.267738	0.306712	0.593827

- Gene_ID: Gene ID
- Transcript_ID: Splicing variant (isoform/transcript)
- Gene_Symbol: Symbol of gene
- Gene_Description: Description of gene
- [Sample Name]_Read_Count: Read count of a sample
- [Sample Name]_FPKM: FPKM normalized value of a sample

5. Differentially Expressed Gene Analysis Results

5.1. Data Analysis Quality Check and Preprocessing

There is a process that sorts differentially expressed gene among samples by read count value of known genes. In preprocessing, there are data quality and similarity checks among samples in case of biological replicates exist.

(Refer to Path: result_RNAseq_excel/DEG_result/Analysis_Result.html)

5.1.1. Sample Information and Analysis Design

Total of 24 samples was used for analysis. For more information of samples and comparison pair, please refer to Sample.Info.txt file.

Index	Sample.ID	Sample.Group
1	NM24_1	NM24
2	NM24_2	NM24
3	NM24_3	NM24
4	AM24_1	AM24
5	AM24_2	AM24
6	AM24_3	AM24
7	NM24_INF_1	NM24_INF
8	NM24_INF_2	NM24_INF
9	NM24_INF_3	NM24_INF
10	AM24_INF_1	AM24_INF
11	AM24_INF_2	AM24_INF
12	AM24_INF_3	AM24_INF
13	NM48_1	NM48
14	NM48_2	NM48
15	NM48_3	NM48
16	AM48_1	AM48
17	AM48_2	AM48
18	AM48_3	AM48
19	NM48_INF_1	NM48_INF
20	NM48_INF_2	NM48_INF
21	NM48_INF_3	NM48_INF
22	AM48_INF_1	AM48_INF
23	AM48_INF_2	AM48_INF

24	AM48_INF_3	AM48_INF
----	------------	----------

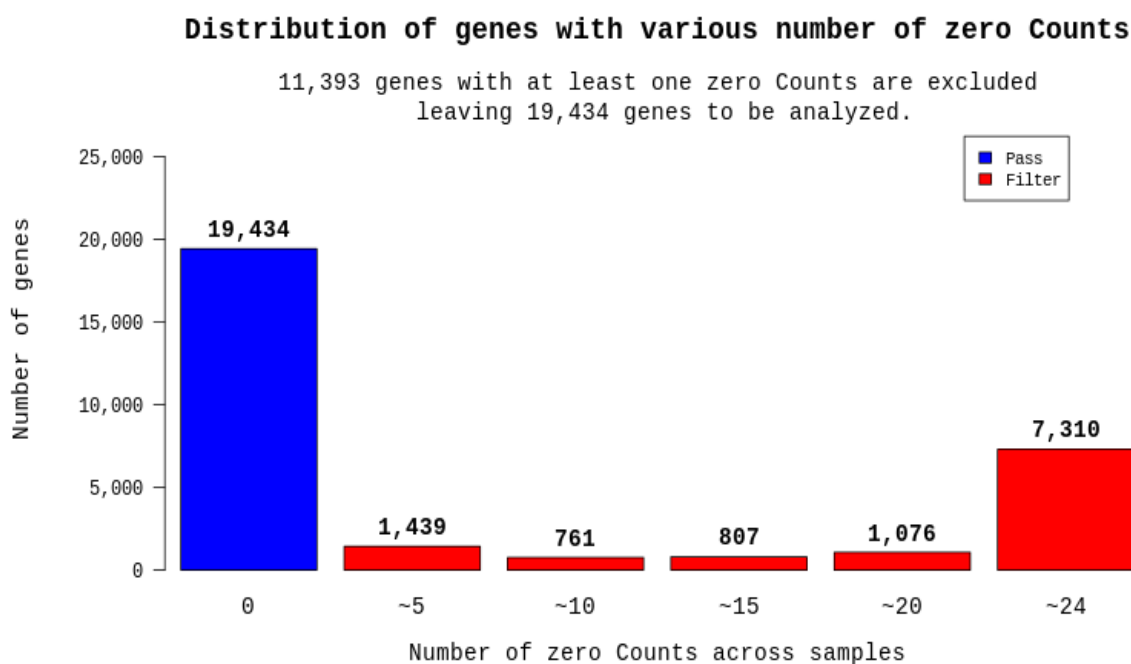
Comparison pair and statistical method for each pair are shown below.

Index	Test vs. Control	Statistical Method
1	AM24 vs. NM24	Fold Change, nbinomWaldTest using DESeq2, Hierarchical Clustering
2	NM24_INF vs. NM24	Fold Change, nbinomWaldTest using DESeq2, Hierarchical Clustering
3	AM24_INF vs. NM24_INF	Fold Change, nbinomWaldTest using DESeq2, Hierarchical Clustering
4	AM24_INF vs. AM24	Fold Change, nbinomWaldTest using DESeq2, Hierarchical Clustering
5	AM48 vs. NM48	Fold Change, nbinomWaldTest using DESeq2, Hierarchical Clustering
6	NM48_INF vs. NM48	Fold Change, nbinomWaldTest using DESeq2, Hierarchical Clustering
7	AM48_INF vs. NM48_INF	Fold Change, nbinomWaldTest using DESeq2, Hierarchical Clustering
8	AM48_INF vs. AM48	Fold Change, nbinomWaldTest using DESeq2, Hierarchical Clustering

5. 1. 2. DATA Quality Check

(Refer to Path: result_RNAseq_excel/DEG_result/Data Quality Check/)

For 24 samples, if more than one read count value was 0, it was not included in the analysis. Therefore, from total of 30,827 genes, 11,393 were excluded and only 19,434 genes were used for statistic analysis.



5. 1. 3. Data Transformation and Normalization

In order to reduce systematic bias, size factors were estimated from the read count data (estimateSizeFactors method).

Using them, the read count data was normalized with Relative Log Expression (RLE) method in DESeq2 R library.

Then, statistical test was performed with the normalized data.

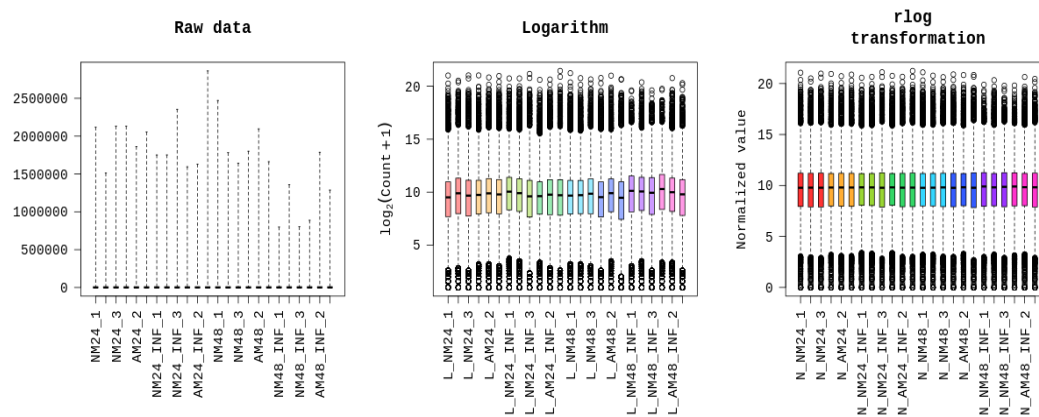
$\log_2(\text{read count}+1)$ and regularized log (rlog) transformed values were used for data visualization. rlog transformation is a method to minimize differences between samples for genes/transcripts in low expression. It transforms count data into \log_2 scale and normalizes them with a library size factor. rlog is robust in the case when the size factors vary widely.

These logarithm figures were used only for visualization.

To proceed a statistical test, RLE normalized count was adopted for negative binomial Wald Test (nbinomWaldTest) in DESeq2.

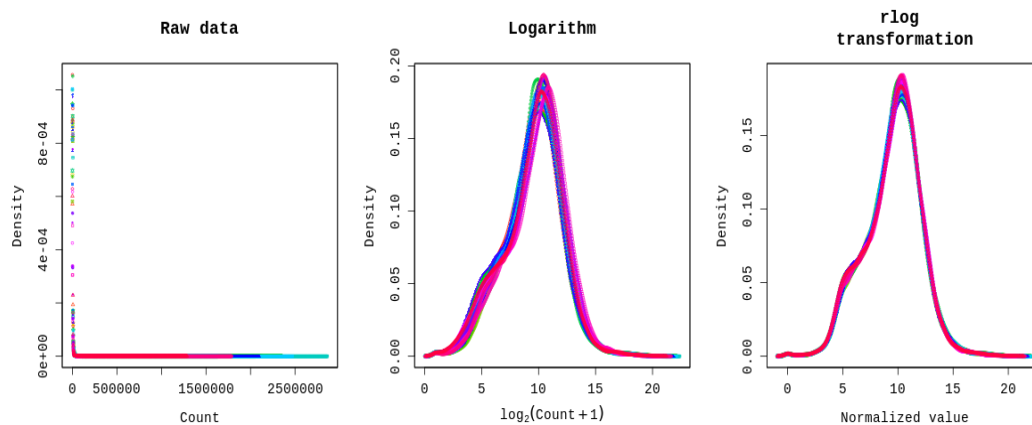
5.1.3.1. Boxplot of Expression Difference between samples.

Below boxplots show the corresponding sample's expression distribution based on percentile (median, 50 percentile, 75 percentile, maximum and minimum) based on raw signal (read count), Log2 transformation of read count+1 and RLE Normalization.



5.1.3.2. Expression Density Plot per sample

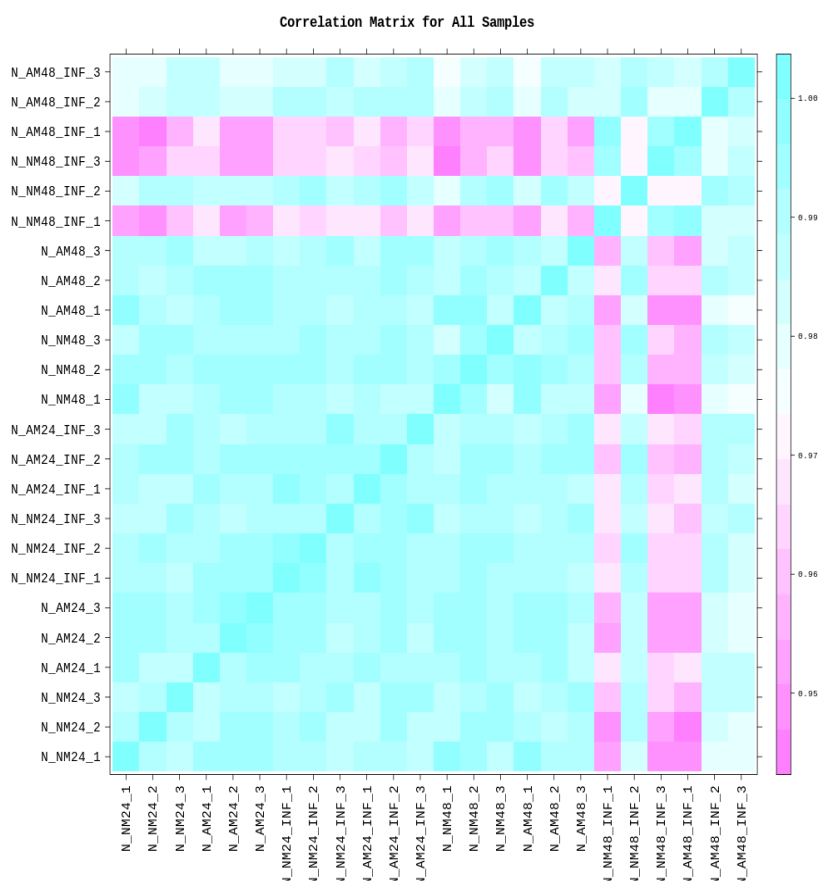
Below density plots show the corresponding samples expression distribution before and after raw signal (read count), Log2 transformation of read count+1 and RLE Normalization.



5. 1. 4. Correlation Analysis between samples

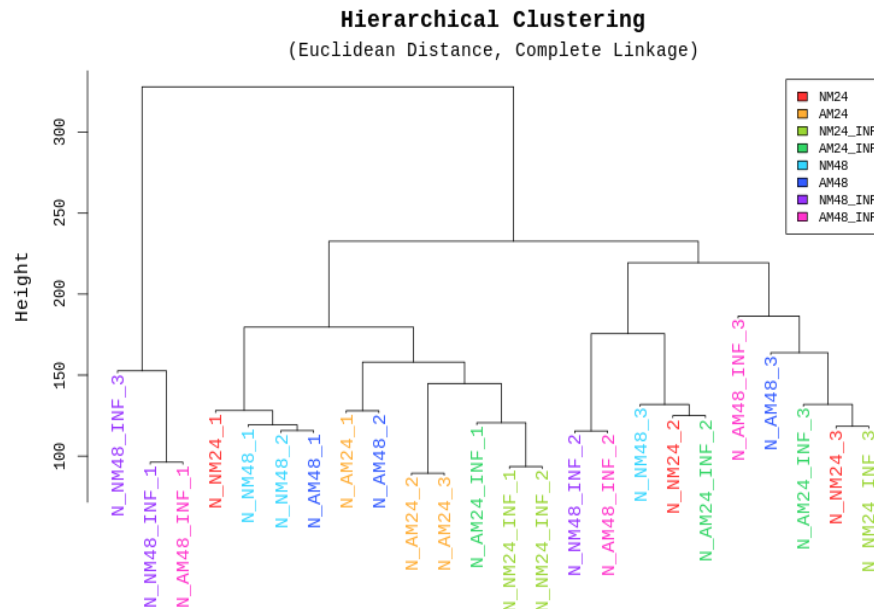
The similarity between samples are obtained through Pearson's coefficient of the normalized value. For range: $-1 \leq r \leq 1$, the closer the value is to 1, the more similar the samples are.

Correlation matrix of all samples is as follows.



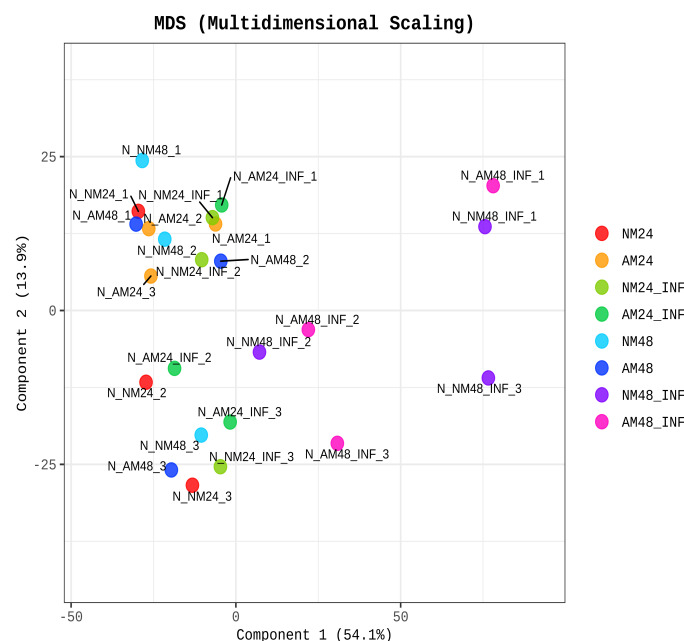
5. 1. 5. Hierarchical Clustering Analysis

Using each sample's normalized value, the high expression similarities were grouped together.
(Distance metric = Euclidean distance, Linkage method= Complete Linkage)



5. 1. 6. Multidimensional Scaling Analysis

Using each sample's normalized value, the similarity between samples is graphically shown in a 2D plot to show the variability of the total data. This allows identification any outlier samples, or similar expression patterns between sample groups.



5. 2. Differentially Expressed Gene Analysis Workflow

Below shows the orders of DEG (Differentially Expressed Genes) analysis.

- 1) the read count value of known genes obtained through -e option of the StringTie were used as the original raw data.

- Raw data

(Refer to Path: result_RNAseq_excel/Expression_profile/StringTie/
Expression_Profile.SL3.0.gene.xlsx)

: 30,827 genes, 24 samples

- 2) During data preprocessing, low quality transcripts are filtered. Afterwards, RLE Normalization are performed.

- Processed data

(Refer to Path: result_RNAseq_excel/DEG_result/data2.xlsx)

: 19,434 genes, 24 samples

- 3) Statistical analysis is performed using Fold Change, nbinomWaldTest using DESeq2 per comparison pair.

The significant results are selected on conditions of $|fc| \geq 2$ & nbinomWaldTest raw p-value < 0.05 .

(data3_*.xlsx contains significant genes which satisfied $|fc| \geq 2$ & nbinomWaldTest raw p-value < 0.05 conditions in more than at least one of total comparison pairs.)

(Refer to Path: result_RNAseq_excel/DEG_result/)

- Significant data (data3_fc2_&_raw.p.xlsx)

: 4,017 genes

- Significant data (data3_AM24_vs_NM24_fc2_&_raw.p.xlsx)

: 239 genes

- Significant data (data3_NM24_INF_vs_NM24_fc2_&_raw.p.xlsx)

: 307 genes

- Significant data (data3_AM24_INF_vs_NM24_INF_fc2_&_raw.p.xlsx)

: 47 genes

- Significant data (data3_AM24_INF_vs_AM24_fc2_&_raw.p.xlsx)

: 353 genes

- Significant data (data3_AM48_vs_NM48_fc2_&_raw.p.xlsx)

: 80 genes

- Significant data (data3_NM48_INF_vs_NM48_fc2_&_raw.p.xlsx)

: 3,335 genes

- Significant data (data3_AM48_INF_vs_NM48_INF_fc2_&_raw.p.xlsx)

: 58 genes

- Significant data (data3_AM48_INF_vs_AM48_fc2_&_raw.p.xlsx)

: 2,057 genes

4) For significant lists, hierarchical clustering analysis is performed to group the similar samples and genes. These results are graphically depicted using heatmap and dendogram.

- Hierarchical Clustering (Euclidean Distance, Complete Linkage)

(Refer to Path: result_RNAseq_excel/DEG_result/Cluster image/)

5) For significant lists, gene-set enrichment analysis was performed based on gene ontology(
<https://biit.cs.ut.ee/gprofiler/>).

Please refer to the GO_stat sheet and the GO_genes sheet of data3 file.

Following result are provided.

- GO_stat
- GO_genes

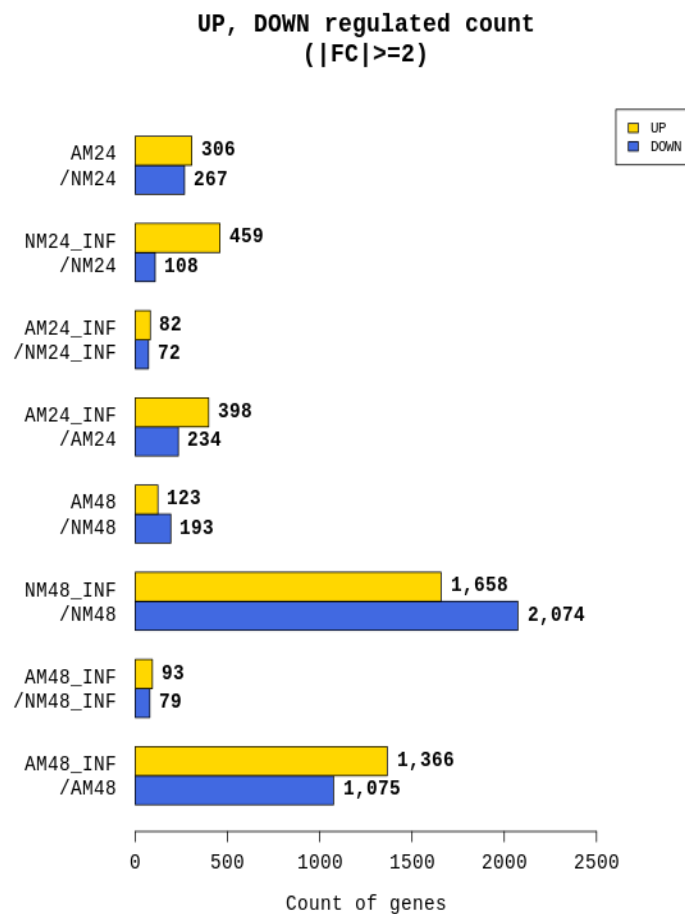
5. 3. Significant Gene Results

(Refer to Path: result_RNAseq_excel/DEG_result/Plots/)

These are DEG result of AM24_vs_NM24 meeting fc2_&_raw.p by example.

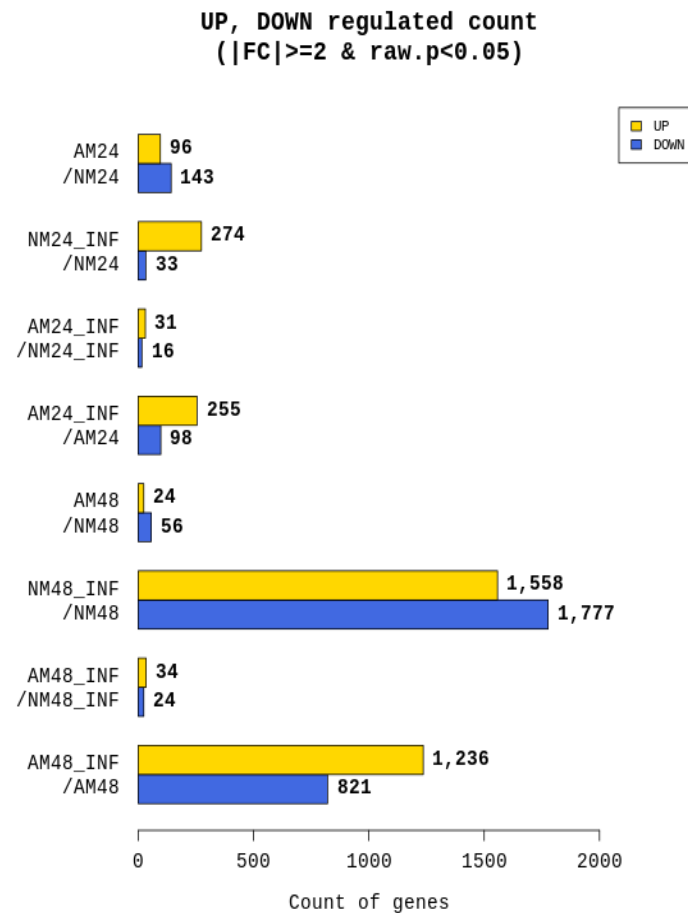
5. 3. 1. Up, Down Regulated Count by Fold Change

Shows number of up and down regulated genes based on fold change of comparison pair.



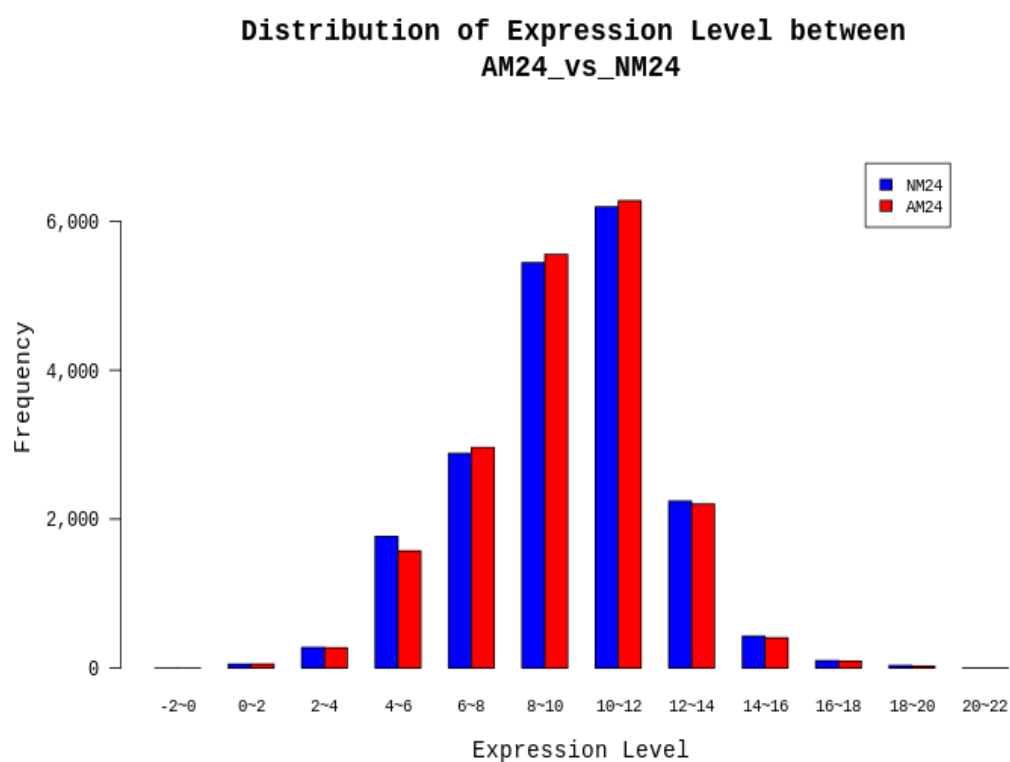
5. 3. 2. Up, Down Regulated Count by Fold Change and p-value

Shows number of up and down regulated genes based on fold change and p-value of comparison pair.



5. 3. 3. Distribution of Expression Level between two groups

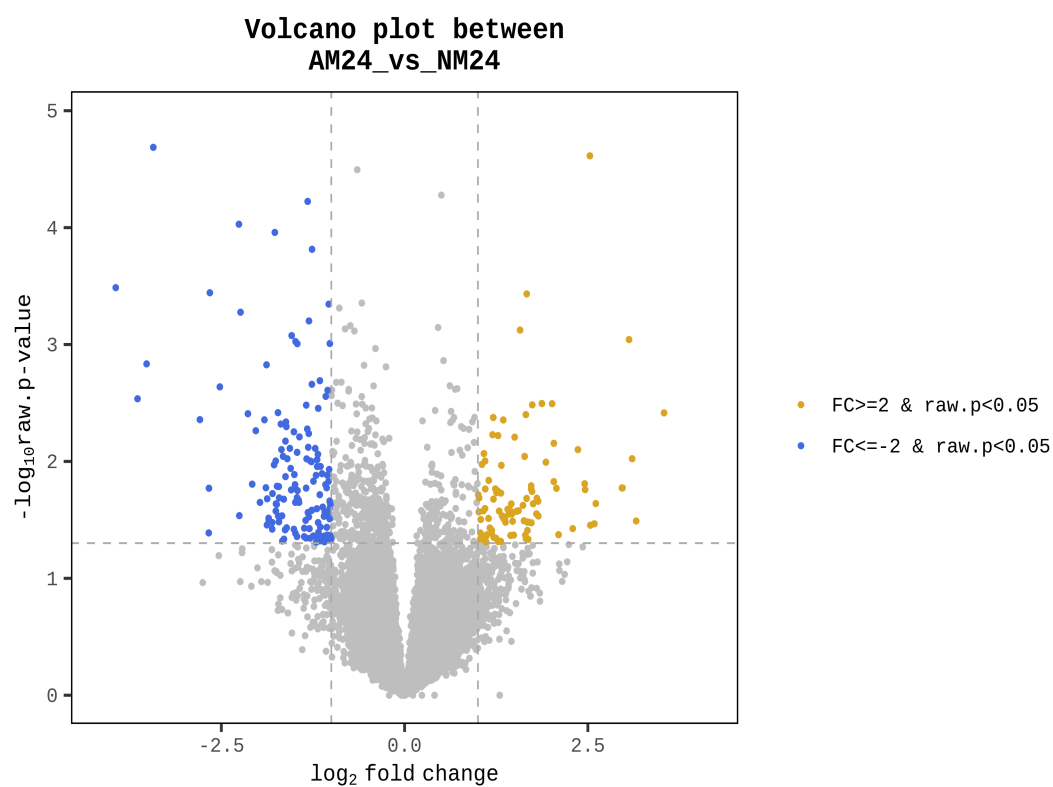
Shows distribution of normalized value of each group for comparison pair.



5. 3. 4. Volcano Plot of Expression Level of two groups.

Log2 fold change and p-value obtained from the comparison between two groups plotted as volcano plot.

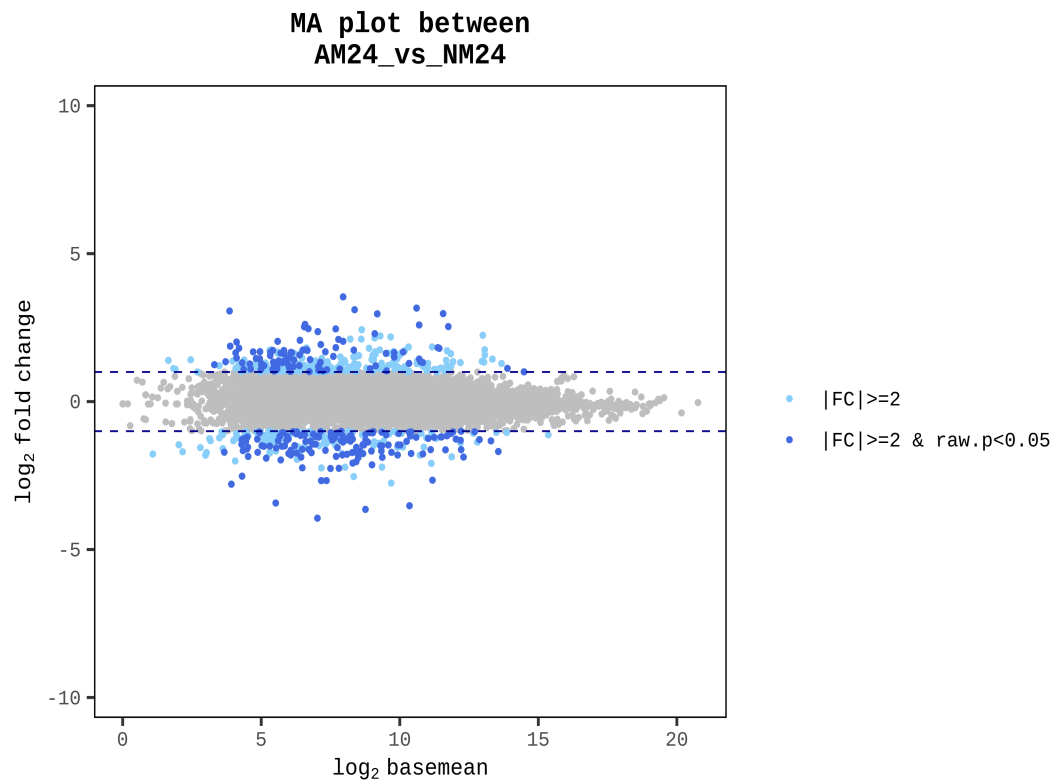
(X-axis: log2 Fold Change, Y-axis: $-\log_{10}$ p-value)



5. 3. 5. MA Plot

In order to confirm the transcripts that show higher expression difference compared to the control according to overall average expression level, MA plot is drawn. (X-axis: mean of normalized counts, Y-axis: log₂ Fold Change).

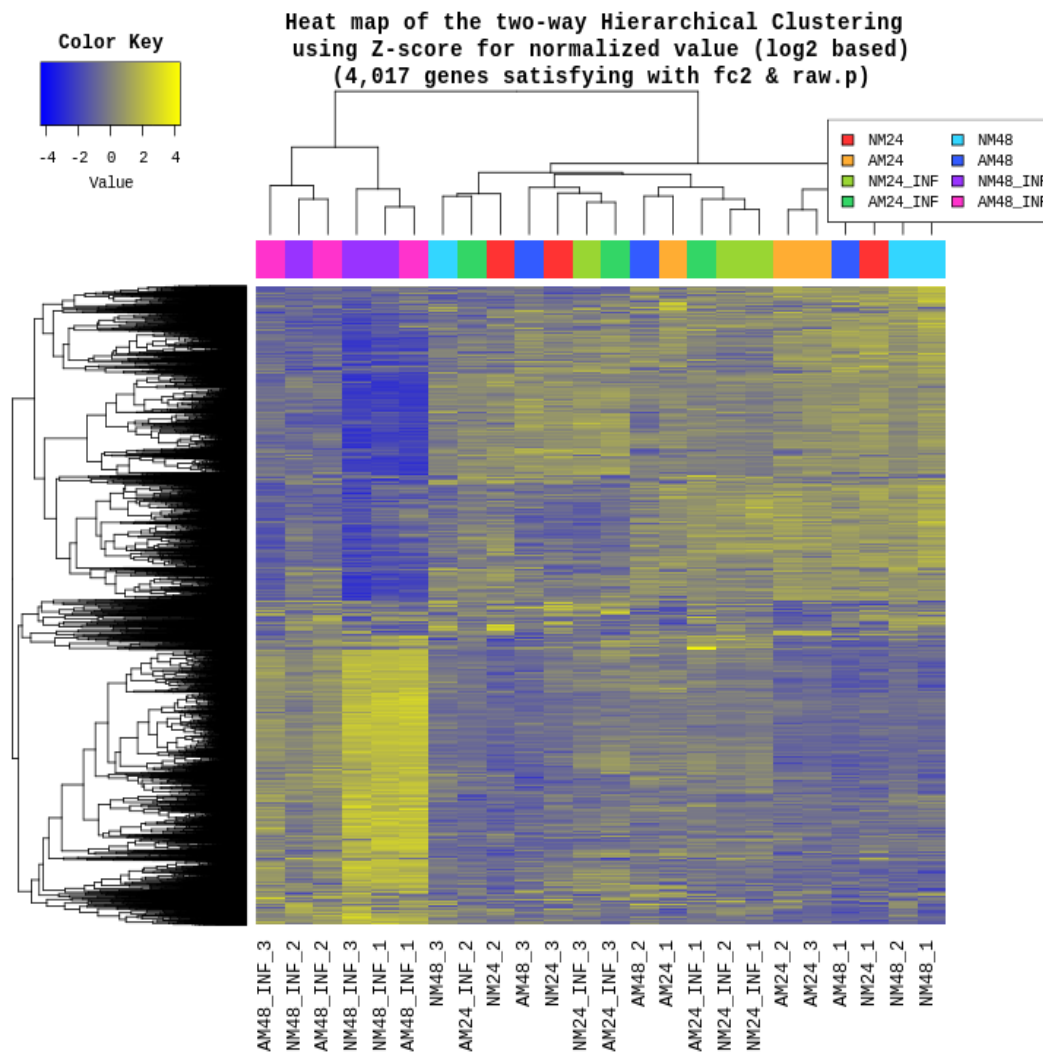
For example, even though fold change might be different by two-fold, the gene with higher mean of normalized counts may be more credible.



5. 3. 6. Hierarchical Clustering Analysis

(Refer to Path: result_RNAseq_excel/DEG_result/Cluster image/)

Heatmap shows result of hierarchical clustering analysis (Euclidean Method, Complete Linkage) which clusters the similarity of genes and samples by expression level (normalized value) from significant list more than at least one of total comparison pairs.



5. 4. GO Enrichment Analysis

(Refer to Path: result_RNAseq_excel/DEG_result/gprofiler)

For Enrichment test which based on Gene Ontology (<http://geneontology.org/>) DB was conducted with significant gene list using g:Profiler tool (<https://biit.cs.ut.ee/gprofiler/>).

The g:Profiler tool performs statistical enrichment analysis to find over-representation of information from Gene Ontology terms, biological pathways, regulatory DNA elements, human disease gene annotations, and protein-protein interaction networks.

Progressing about 3 categories of GO. The gene or gene product, molecule associated with GO ID was summarized by parsing the ontology file and the annotation file (multispecies annotation provided by Uniprot, or the annotation provided by each type reference DB for the GO consortium) for the GO graph structure.

- Link for the ontology documentation: <http://geneontology.org/page/ontology-documentation>
- Link for the ontology files: <http://geneontology.org/page/download-ontology>
- Link for the annotation files: <http://geneontology.org/page/download-annotations>

Enrichment test result was summarized at each sheet of DEG result(data3-*.xlsx file) by 2 forms below.

- GO_stat
- GO_genes

5. 4. 1. GO_stat Sheet

The result of associated gene and test stat was summarized by term_id. The significance of specific term_id in enrichment test with DEG set was summarized.

source	term_id	term_name	adjusted_p_value	term_size	query_size	intersection_size	effective_domain_size	intersections
GO:CC	GO:0022626	cytosolic ribosome	2.72198E-17	115	1921	50	18797	6134, 6206, 6155, 6204, 6168, 200916,
GO:BP	GO:0006614	SRP-dependent cotranslational protein targeting to membrane	3.60328E-15	96	1824	44	17816	6134, 6206, 6155, 6204, 6168, 6747, 61
GO:MF	GO:0003735	structural constituent of ribosome	1.32911E-14	170	1860	59	18098	6134, 6206, 6155, 6204, 6168, 200916,
GO:BP	GO:0006613	cotranslational protein targeting to membrane	2.03613E-14	101	1824	44	17816	6134, 6206, 6155, 6204, 6168, 6747, 61
GO:MF	GO:0005198	structural molecule activity	4.45523E-14	739	1860	151	18098	6134, 6206, 127294, 4586, 301, 3887, 6
GO:BP	GO:0045047	protein targeting to ER	7.18306E-14	109	1824	45	17816	6134, 6206, 6155, 6204, 6168, 6747, 61
GO:CC	GO:0044391	ribosomal subunit	2.36014E-13	195	1921	61	18797	6134, 6206, 6155, 6204, 6168, 200916,
GO:BP	GO:0072599	establishment of protein localization to endoplasmic reticulum	2.82077E-13	113	1824	45	17816	6134, 6206, 6155, 6204, 6168, 6747, 61
GO:BP	GO:0070972	protein localization to endoplasmic reticulum	4.06119E-11	137	1824	47	17816	6134, 6206, 6155, 6204, 6168, 6747, 61
GO:CC	GO:0005840	ribosome	1.34069E-10	246	1921	65	18797	6134, 6206, 6155, 6204, 6168, 200916,
GO:CC	GO:0022625	cytosolic large ribosomal subunit	1.69728E-10	64	1921	29	18797	6134, 6155, 6168, 200916, 6167, 6161,
GO:BP	GO:000184	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	7.11348E-10	122	1824	42	17816	6134, 6206, 6155, 6204, 6168, 6167, 61
GO:CC	GO:0044459	plasma membrane part	1.34094E-09	2879	1921	400	18797	165829, 10326, 6405, 4283, 8322, 5743
GO:CC	GO:0071944	cell periphery	1.8891E-09	5662	1921	709	18797	829, 165829, 10326, 23256, 6405, 4283
GO:CC	GO:0005886	plasma membrane	5.37824E-09	5539	1921	692	18797	165829, 10326, 23256, 6405, 4283, 505
GO:CC	GO:0044444	cytoplasmic part	5.37824E-09	9685	1921	1125	18797	6134, 829, 84532, 10326, 5332, 23256,
GO:CC	GO:0005737	cytoplasm	5.47219E-09	11534	1921	1309	18797	6134, 829, 84532, 10326, 5332, 23256,
GO:BP	GO:0009888	tissue development	5.79564E-09	2068	1824	305	17816	6405, 5054, 8322, 5743, 144165, 12729
GO:BP	GO:0006612	protein targeting to membrane	5.95069E-09	195	1824	54	17816	6134, 6206, 6155, 6204, 6168, 6747, 51
GO:BP	GO:0051179	localization	1.23607E-08	6751	1824	824	17816	6134, 829, 10326, 10734, 23256, 6405,
GO:CC	GO:1903561	extracellular vesicle	1.65132E-08	2165	1921	309	18797	829, 5054, 10103, 2098, 9518, 4151, 41
GO:CC	GO:0043230	extracellular organelle	1.66899E-08	2167	1921	309	18797	829, 5054, 10103, 2098, 9518, 4151, 41
GO:CC	GO:0044445	cytosolic part	2.71585E-08	252	1921	60	18797	6134, 6206, 6155, 6204, 6168, 338321,
GO:BP	GO:0032501	multicellular organismal process	3.22915E-08	7718	1824	922	17816	6134, 829, 6405, 5670, 5054, 7079, 832

- source: Code for the data source. Ex> GO:BP | GO:CC | GO:MF ...
- term_id: ID for the enriched term/functional category
- term_name: Readable name for the enriched term
- adjusted_p_value: Adjusted p-value by FDR
- query_size: The number of unique DEG that are annotated to the data source (the functional category).
- intersection_size: The number of unique DEG that are annotated to the term_id
- term_size: The number of genes of species that are annotated to the term_id.
- effective_domain_size: The number of genes of species that are annotated to the data source (the functional category).
- intersections: list of unique DEG that are annotated to the term_id

5. 4. 2. GO_genes Sheet

The result of associated term_id and DEG analysis result was summarized based on Gene. term_id which associated with specific gene was summarized with stat such as fold change, p-value, volume, normalized value.

source	term_id	term_name	adjusted_p_value	intersection_size	Gene_ID	Transcript_ID	Gene_Symbol	test/control.fc	test/control.logCPM	test/control.raw.pval	test/control.bh.pval	N_control_1	N_control_2	N_test_1	N_test_2	
GO:CC	GO:0044444	cytoplasmic part	5.37824E-09	1125	9	NM_000662.NN	NAT1	2.593577	0.965259	1.66575E-06	1.40133E-05	1.167645	0.902212	1.879864	1.926688	
GO:CC	GO:0005737	cytoplasm	5.47219E-09	1309	9	NM_000662.NN	NAT1	2.593577	0.965259	1.66575E-06	1.40133E-05	1.167645	0.902212	1.879864	1.926688	
GO:BP	GO:0070887	cellular response to ch	6.97255E-05	417	9	NM_000662.NN	NAT1	2.593577	0.965259	1.66575E-06	1.40133E-05	1.167645	0.902212	1.879864	1.926688	
GO:BP	GO:0050896	response to stimulus	0.000405505	1045	9	NM_000662.NN	NAT1	2.593577	0.965259	1.66575E-06	1.40133E-05	1.167645	0.902212	1.879864	1.926688	
GO:CC	GO:0005829	cytosol	0.078450245	563	9	NM_000662.NN	NAT1	2.593577	0.965259	1.66575E-06	1.40133E-05	1.167645	0.902212	1.879864	1.926688	
GO:CC	GO:0005622	intracellular	0.110987379	1522	9	NM_000662.NN	NAT1	2.593577	0.965259	1.66575E-06	1.40133E-05	1.167645	0.902212	1.879864	1.926688	
GO:MF	GO:0040600	arylamine N-acetyltras	0.573292063	1	9	NM_000662.NN	NAT1	2.593577	0.965259	1.66575E-06	1.40133E-05	1.167645	0.902212	1.879864	1.926688	
GO:CC	GO:0005575	cellular_component		1	1921	9	NM_000662.NN	NAT1	2.593577	0.965259	1.66575E-06	1.40133E-05	1.167645	0.902212	1.879864	1.926688
GO:BP	GO:0008150	biological_process		1	1824	9	NM_000662.NN	NAT1	2.593577	0.965259	1.66575E-06	1.40133E-05	1.167645	0.902212	1.879864	1.926688
GO:CC	GO:0044459	plasma membrane pa	1.34094E-09	400	24	NM_000350	ABCA4	-8.936138	3.797432	1.4729E-62	6.89976E-60	4.626902	4.764929	1.879864	1.961991	
GO:CC	GO:0071944	cell periphery	1.88911E-09	709	24	NM_000350	ABCA4	-8.936138	3.797432	1.4729E-62	6.89976E-60	4.626902	4.764929	1.879864	1.961991	
GO:CC	GO:0016020	membrane	0.000332978	1085	24	NM_000350	ABCA4	-8.936138	3.797432	1.4729E-62	6.89976E-60	4.626902	4.764929	1.879864	1.961991	
GO:CC	GO:0097458	neuron part	0.000244106	234	24	NM_000350	ABCA4	-8.936138	3.797432	1.4729E-62	6.89976E-60	4.626902	4.764929	1.879864	1.961991	
GO:CC	GO:0042995	cell projection	0.000353388	283	24	NM_000350	ABCA4	-8.936138	3.797432	1.4729E-62	6.89976E-60	4.626902	4.764929	1.879864	1.961991	
GO:CC	GO:0044425	membrane part	0.000390502	813	24	NM_000350	ABCA4	-8.936138	3.797432	1.4729E-62	6.89976E-60	4.626902	4.764929	1.879864	1.961991	
GO:BP	GO:0050896	response to stimulus	0.000405505	1045	24	NM_000350	ABCA4	-8.936138	3.797432	1.4729E-62	6.89976E-60	4.626902	4.764929	1.879864	1.961991	
GO:BP	GO:0051606	detection of stimulus		1	35	24	NM_000350	ABCA4	-8.936138	3.797432	1.4729E-62	6.89976E-60	4.626902	4.764929	1.879864	1.961991
GO:BP	GO:0008150	biological_process		1	1824	24	NM_000350	ABCA4	-8.936138	3.797432	1.4729E-62	6.89976E-60	4.626902	4.764929	1.879864	1.961991
GO:CC	GO:0044444	cytoplasmic part	5.37824E-09	1125	34	NM_000016.NN	ACADM	2.326451	4.229202	9.93422E-14	2.78049E-12	3.715210	3.505224	4.754088	4.772040	
GO:CC	GO:0005737	cytoplasm	5.47219E-09	1309	34	NM_000016.NN	ACADM	2.326451	4.229202	9.93422E-14	2.78049E-12	3.715210	3.505224	4.754088	4.772040	
GO:BP	GO:0009888	tissue development	5.79564E-09	305	34	NM_000016.NN	ACADM	2.326451	4.229202	9.93422E-14	2.78049E-12	3.715210	3.505224	4.754088	4.772040	
GO:BP	GO:0032501	multicellular organism	3.22915E-08	922	34	NM_000016.NN	ACADM	2.326451	4.229202	9.93422E-14	2.78049E-12	3.715210	3.505224	4.754088	4.772040	
GO:BP	GO:0048731	system development	3.55854E-08	626	34	NM_000016.NN	ACADM	2.326451	4.229202	9.93422E-14	2.78049E-12	3.715210	3.505224	4.754088	4.772040	
GO:BP	GO:0048513	animal organ develop	3.78565E-08	478	34	NM_000016.NN	ACADM	2.326451	4.229202	9.93422E-14	2.78049E-12	3.715210	3.505224	4.754088	4.772040	

- source: Code for the data source. Ex> GO:BP | GO:CC | GO:MF ...
- term_id: ID for the enriched term/functional category
- term_name: Readable name for the enriched term
- adjusted_p_value: Adjusted p-value by FDR
- intersection_size: The number of unique DEG that are annotated to the term_id

data3.GO_*.gprofiler.png: Top 20 terms of Gene Ontology Enrichment Analysis result were described by dot plot.

(Plotting based on GO_stat)

data3.GO_*.gprofiler.sizefilt.png: After term_size filtering (min=10, max=500), top 20 terms of Gene Ontology Enrichment Analysis result were described by dot plot.

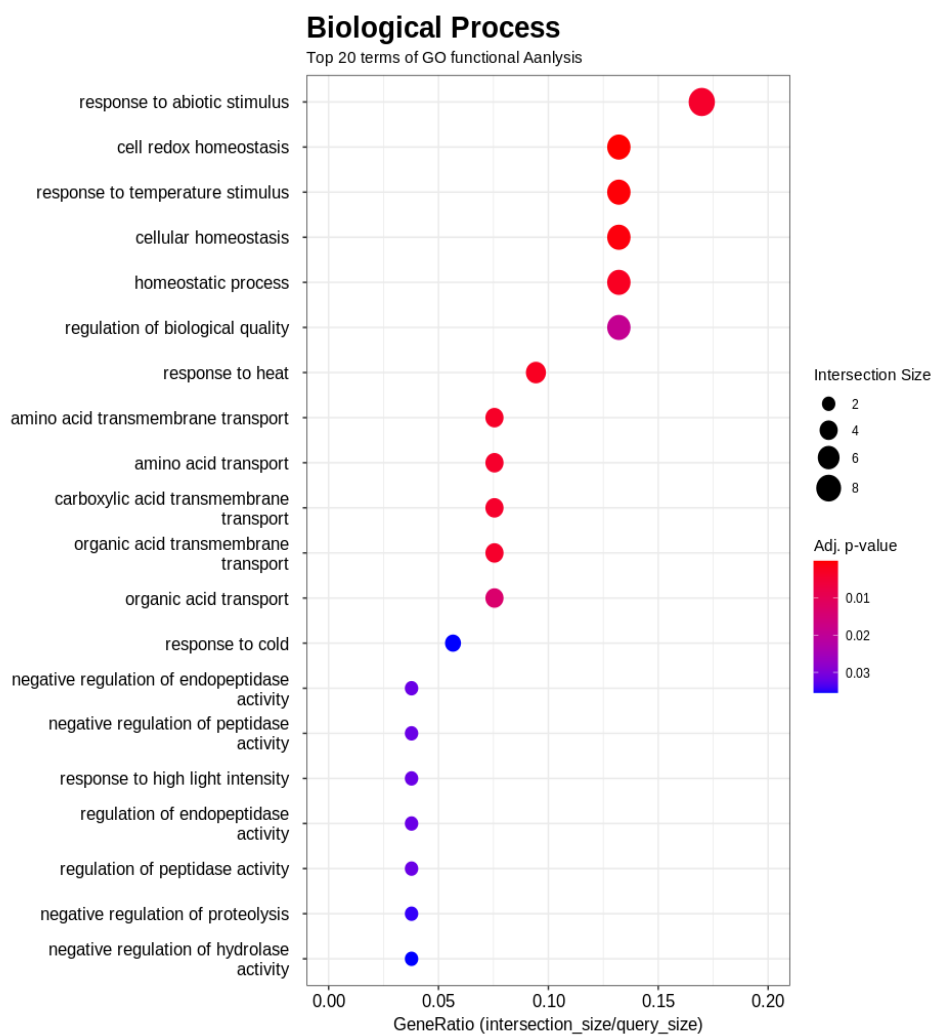
(Plotting based on GO_stat. Please refer to ./gprofiler/data3*.GO/folder.)

- term_size filtering: The GO Terms that are very large or small do not contribute to interpretability of results, and their statistical significance can be inflated when using certain statistical enrichment methods (e.g., Hypergeometric test).

- GeneRatio: GeneRatio is calculated as the ratio of intersection_size and query_size.

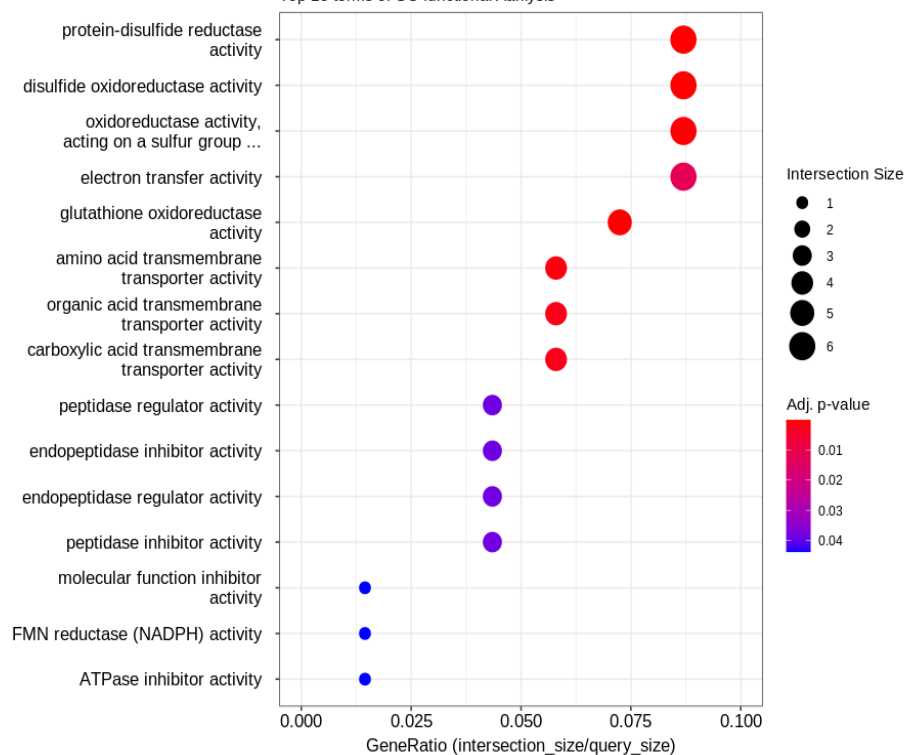
The dot plot below shows the results of the enrichment analysis based on Gene Ontology DB for significant genes.

These dot plots are examples for data3.GO_*.gprofiler.png (without term_size filtering).



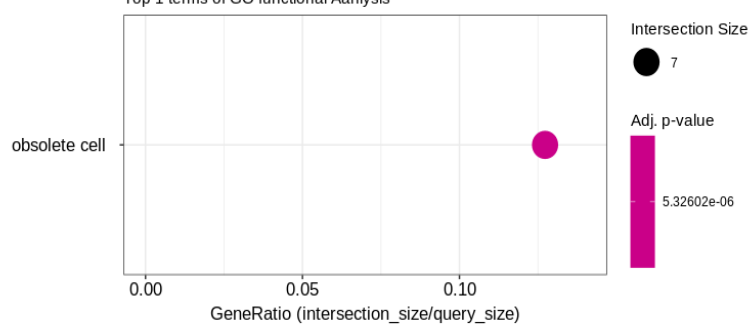
Molecular Function

Top 15 terms of GO functional Analysis



Cellular Component

Top 1 terms of GO functional Analysis



6. Data Download Information

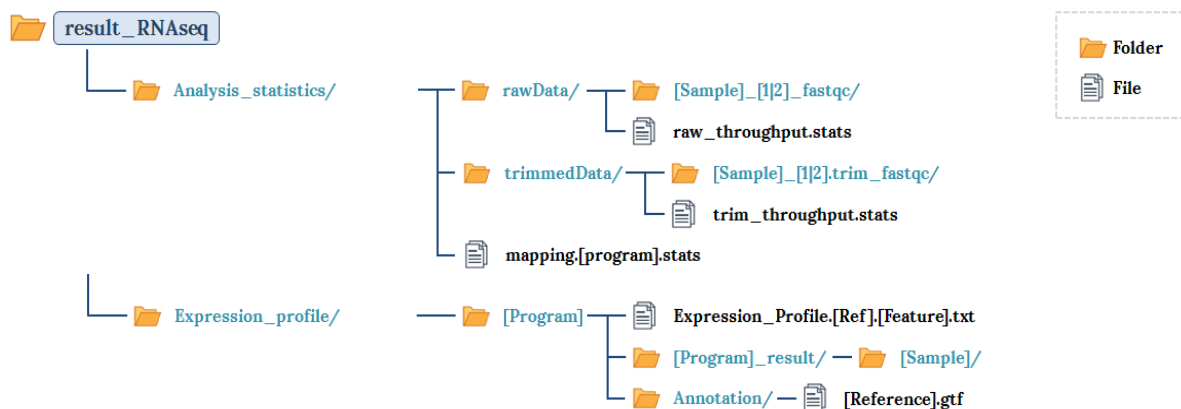
6.1. Raw Data

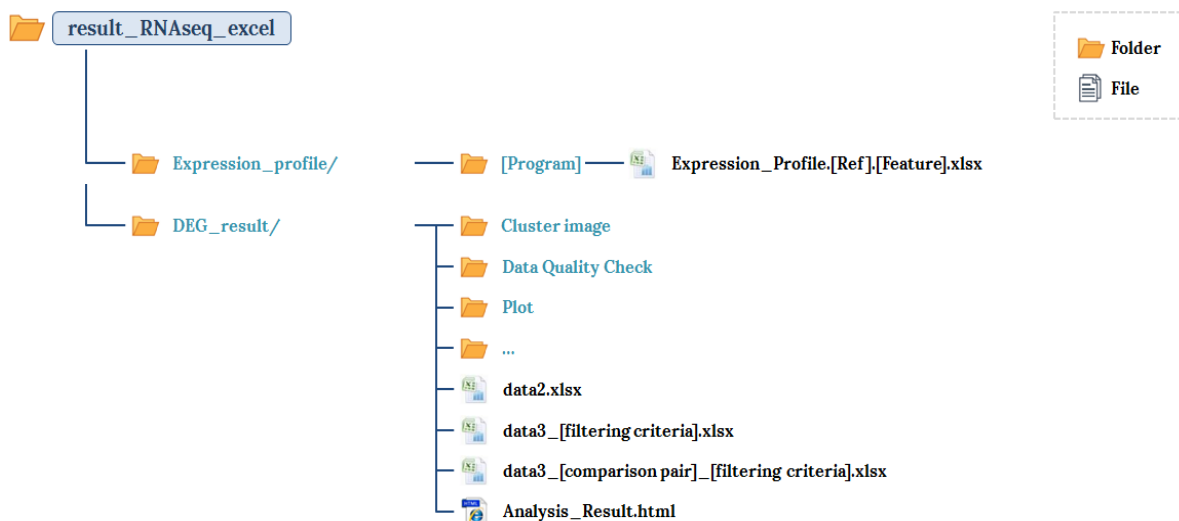
Raw data is the FASTQ file that isn't trimmed adapter sequence.


File name	File size	md5sum
AM24_1_1.fastq.gz	2.44G	dc17f87479028b77fb22a1ea35526652
AM24_1_2.fastq.gz	2.51G	2ccfff98016de87082485479eff70e59
AM24_2_1.fastq.gz	2.69G	551ea5217302fe88a5b1716553ccb05e
AM24_2_2.fastq.gz	2.77G	36bceba38ad634b9958737c42d1bbd57
AM24_3_1.fastq.gz	2.6G	c3fdd78d36e8b93dd064963f241cdf98
AM24_3_2.fastq.gz	2.69G	de9aed08eb321161098d61dlb40393ec
AM24_INF_1_1.fastq.gz	2.22G	be432c9cd59a25432274bf0ddece8b01
AM24_INF_1_2.fastq.gz	2.25G	5c5b47b31a051c998771f83653bf5759
AM24_INF_2_1.fastq.gz	2.48G	411691376c9be1e5a2e2334d6f9d50ee
AM24_INF_2_2.fastq.gz	2.55G	7f83655e4edaadd7a8101d48c81c943f
AM24_INF_3_1.fastq.gz	2.81G	6f58abe9b92b7a24b444d0d6095dc280
AM24_INF_3_2.fastq.gz	2.88G	c92714bb9ddeefcf9471f984e4264876
AM48_1_1.fastq.gz	2.2G	b77fb59acd2ce9eaa40d28c2f654123d
AM48_1_2.fastq.gz	2.3G	b6731b8b9a64385b892cacd77f2cead8
AM48_2_1.fastq.gz	2.56G	c1ff45f477d0fe4859dbfc513938eb5b
AM48_2_2.fastq.gz	2.6G	4f1908920a9a192587e283cbfa1a21b3
AM48_3_1.fastq.gz	2.32G	09fb20c454b5510268ab87170000fd50
AM48_3_2.fastq.gz	2.39G	0d0078230fb978f5e06b3fc8d8c017ec
AM48_INF_1_1.fastq.gz	2.66G	f692ed0dbabca0a087f5ec233849f515
AM48_INF_1_2.fastq.gz	2.72G	27648c107bb6c4eeecbb103b6e7f8af3
AM48_INF_2_1.fastq.gz	2.58G	64620153cd3871e7dd73d8a74f1214a7
AM48_INF_2_2.fastq.gz	2.63G	70a3f0a6d937026145df0640d4857686
AM48_INF_3_1.fastq.gz	2.3G	806765ce69f06688892c701175f54ba8
AM48_INF_3_2.fastq.gz	2.36G	2fe1069944a44e447af20ae51e0d673d
NM24_1_1.fastq.gz	2.39G	04c7fd7ba7d82ae200518d746cb45bb1
NM24_1_2.fastq.gz	2.49G	d56e449cd0e096eb8943dfc36752208c
NM24_2_1.fastq.gz	2.69G	cd3b3beb86c7548c1c3f5dcf759d7938
NM24_2_2.fastq.gz	2.75G	842e870ba0cb93c9a558314f88017ce2
NM24_3_1.fastq.gz	2.64G	87a11abc87dfd6f7915a37fc75c97f35
NM24_3_2.fastq.gz	2.75G	7f71393f15540a60160c2cb84fe35431
NM24_INF_1_1.fastq.gz	2.77G	961353d3de771956b57e1e259d206302
NM24_INF_1_2.fastq.gz	2.84G	2f9aee8e9f4c90ea0ca65fa88311cc4a

NM24_INF_2_1.fastq.gz	2.68G	e00bfb35d9deca47f285173322f1834
NM24_INF_2_2.fastq.gz	2.77G	08fb928498ed929ff5596916d843af44
NM24_INF_3_1.fastq.gz	2.64G	d35e58ed8f8dfa6923d83ce3e9698c99
NM24_INF_3_2.fastq.gz	2.72G	573cfc720a93dbb2565d266068c40cfa
NM48_1_1.fastq.gz	2.58G	e3e0c957f6c9bd0dd0bc7a1620a2422b
NM48_1_2.fastq.gz	2.67G	1a5f29308ed3295b37493959a740e12a
NM48_2_1.fastq.gz	2.38G	82b0d9dc46639c998c763048c072941a
NM48_2_2.fastq.gz	2.44G	89be38d58cb97975e75cdcd54202786b
NM48_3_1.fastq.gz	2.56G	33eb78f4a10a59a693d4e1891afed52f
NM48_3_2.fastq.gz	2.63G	ee408e7b4a5b69a460a913243e51944e
NM48_INF_1_1.fastq.gz	2.46G	0d19b6bb7a89a9ca8716c495dfe8ebdf
NM48_INF_1_2.fastq.gz	2.53G	9887d7913a2c7a99354c81a3a1a42a5d
NM48_INF_2_1.fastq.gz	2.58G	0416ceff62be940d1e7de892e550f076
NM48_INF_2_2.fastq.gz	2.68G	dad8ec155e9b7f05eef77fe7b793bf27
NM48_INF_3_1.fastq.gz	2.22G	f0c4b6724be2dc9699e9e3a5eb09ba27
NM48_INF_3_2.fastq.gz	2.3G	cef30bf1dee0f7cb0641cf9b3a9d71b3

- fastq.gz : This is a zip file of raw data used in analysis.
- md5sum : In order to verify the integrity of files, md5sum is used. If the values of md5sum are the same, there is no forgery, modification or omission.





 Your data will be retained in our server for 3 months.
Should you wish to extend the retention period, please contact us.

7. Appendix

7.1. Phred Quality Score Chart

Phred quality score numerically express the accuracy of each nucleotide. Higher Q number signifies higher accuracy. For example, if Phred assigns a quality score of 30 to a base, the chances of having base call error are 1 in 1000.

Quality of phred score	Probability of incorrect base call	Base call accuracy	Characters
10	1 in 10	90%	!"#\$%&'()*+,-./0123456789:;h=i?
20	1 in 100	99%	!,"#\$%&'()*+,-./0123456789:;h=i?@
30	1 in 1000	99.9%	!,"#\$%&'()*+,-./0123456789:;h=i?@ABCDEFGHIJ
40	1 in 10000	99.99%	!,"#\$%&'()*+,-./0123456789:;h=i?@ABCDEFGHIJ

Phred Quality Score Q is calculated with $-10\log_{10}P$, where P is probability of erroneous base call.

7. 2. Programs used in Analysis

7. 2. 1. FastQC v0.11.7

LINK <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

FastQC is a program that performs quality check on the raw sequences before analysis to make sure data integrity. The main function is importing BAM, SAM, FastQ files and providing quick overview on which section has problems. It provides such results as graphs and tables in html files.

7. 2. 2. Trimmomatic 0.38

LINK <http://www.usadellab.org/cms/?page=trimmomatic>

Trimmomatic is a program that performs trimming depending on various parameters on illumina paired-end or single-end.

- ILLUMINACLIP: Cut adapter and other illumina-specific sequences from the read.
- SLIDINGWINDOW: Perform a sliding window trimming, cutting once the average quality within the window falls below a threshold.
- LEADING: Cut bases off the start of a read, if below a threshold quality.
- TRAILING: Cut bases off the end of a read, if below a threshold quality.
- CROP: Cut the read to a specified length.
- HEADCROP: Cut the specified number of bases from the start of the read.
- MINLEN: Drop the read if it is below a specified length.
- TOPHRED33: Change quality score to phred33.
- TOPHRED64: Change quality score to phred64.

7. 2. 3. HISAT2 version 2.1.0, Bowtie2 2.3.4.1

LINK <https://ccb.jhu.edu/software/hisat2/index.shtml>

HISAT2 is a fast and sensitive alignment program for mapping next-generation sequencing reads to genomes. Its first implementation based on an extension of BWT for graphs, designed a graph FM index (GFM). In addition to using one global GFM index, HISAT2 uses a large set of small GFM indexes that collectively cover the whole genome (each index representing a genomic region of 56 Kbp, with 55,000 indexes needed to cover the human population). These small indexes (called local indexes), combined with several alignment strategies, enable rapid and accurate alignment of sequencing reads. This new indexing scheme is called a Hierarchical Graph FM index (HGFM).

7. 2. 4. StringTie version 2.1.3b

LINK <https://ccb.jhu.edu/software/stringtie/>

StringTie is a fast and highly efficient assembler of RNA-Seq alignments into potential transcripts. It uses a novel network flow algorithm as well as an optional de novo assembly step to assemble and quantitate full-length transcripts representing multiple splice variants for each gene locus.

7. 3. References

1. BOLGER, Anthony M.; LOHSE, Marc; USADEL, Bjoern. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 2014, btu170.
2. KIM, Daehwan; LANGMEAD, Ben; SALZBERG, Steven L. HISAT: a fast spliced aligner with low memory requirements. *Nature methods*, 2015, 12.4: 357-360.
3. LI, Heng, et al. The sequence alignment/map format and SAMtools. *Bioinformatics*, 2009, 25.16: 2078-2079.
4. PERTEA, Mihaela, et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature biotechnology*, 2015, 33.3: 290-295.
5. PERTEA, Mihaela, et al. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nature Protocols*, 2016, 11.9: 1650-1667.
6. RAUDVERE, Uku, et al. g: Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic acids research*, 2019.



HEADQUARTER

Macrogen, Inc.

Laboratory, IT and Business Headquarter & Support Center

[08511] 1001, 10F, 254, Beotkkot-ro,
Geumcheon-gu, Seoul, Republic of Korea
(Gasan-dong, World Meridian 1)

Tel: +82-2-2180-7000

Email1: ngs@macrogen.com(Overseas)

Email2: ngskr@macrogen.com

(Republic of Korea)

Web: www.macrogen.com

LIMS: dna.macrogen.com

SUBSIDIARY

Macrogen Europe

Laboratory, Business & Support Center

Meibergdreef 57, 1105 BA, Amsterdam,
the Netherlands

Tel: +31-20-333-7563

Email: ngs@macrogen.eu

Psomagen (Macrogen USA)

Laboratory, Business & Support Center

1330 Piccard Drive, Suite 103, Rockville,
MD 20850, United States

Tel: +1-301-251-1007

Email: inquiry@psomagen.com

Macrogen Singapore

Laboratory, Business & Support Center

3 Biopolis Drive #05-18, Synapse,
Singapore 138623

Tel: +65-6339-0927

Email: info-sg@macrogen.com

Macrogen Japan

Laboratory, Business & Support Center

16F Time24 Building, 2-4-32 Aomi,
Koto-ku, Tokyo 135-0064 JAPAN

Tel: +81-3-5962-1124

Email: ngs@macrogen-japan.co.jp

BRANCH

Macrogen Spain

Laboratory, Business & Support Center

Av. Sur del Aeropuerto de Barajas,
28. Office B-2, 28042 Madrid, Spain

Tel: +34-911-138-378

Email: info-spain@macrogen.com